鲁棒平均奖励马尔可夫决策过程

Yue Wang, [1] **Alvaro Velasquez,** [2] **George Atia,** [3] **Ashley Prater-Bennette,** [4] **Shaofeng Zou** [1]

[1] 布法罗大学，纽约州立大学[2] 科罗拉多大学博尔德分校[3] 中佛罗里达大学[4] 美国空军研究实验室

ywang294@buffalo.com， alvaro.velasquez@colorado.edu，george.atia@ucf.edu，
ashley.prater-bennette@us.af.mil， szou3@buffalo.edu

arXiv:2301.00858v2 [cs.LG] 1 Mar 2023

摘要

在鲁棒马尔可夫决策过程中，针对转移核的不确定性，需要在马尔可夫决策过程的不确定集上寻找使最坏情况性能最优的策略。尽管大量文献聚焦于折扣马尔可夫决策过程，鲁棒平均奖励马尔可夫决策过程仍鲜有研究。本文聚焦于鲁棒平均奖励马尔可夫决策过程，其目标是在一个不确定集上寻找使最坏情况平均奖励最优的策略。我们首先采用一种利用折扣马尔可夫决策过程近似平均奖励马尔可夫决策过程的方法。我们证明，随着折扣因子 γ 趋近于 1，鲁棒折扣价值函数收敛于鲁棒平均奖励；此外，当 γ 足够大时，鲁棒折扣马尔可夫决策过程的任一最优策略也是鲁棒平均奖励马尔可夫决策过程的最优策略。我们进一步设计了一种鲁棒动态规划方法，并从理论上刻画其收敛至最优解的性质。随后，我们不借助折扣马尔可夫决策过程作为中间步骤，直接研究鲁棒平均奖励马尔可夫决策过程。我们推导出鲁棒平均奖励马尔可夫决策过程的鲁棒贝尔曼方程，证明可以由其解导出最优策略，并进一步设计了一种鲁棒相对价值迭代算法，该算法能够以可证明的方式求得该方程的解，或等价地，求得最优鲁棒策略。

引言

马尔可夫决策过程（MDP）是用于随机环境中序贯决策的有效数学工具（Derman 1970；Puterman 1994）。求解马尔可夫决策过程问题，意味着要找到一个依据给定准则使累积奖励最大化的最优策略。然而，在实践中，由于环境非平稳、建模误差、外生扰动、部分可观测性以及对抗攻击等多种因素，所假设的马尔可夫决策过程模型可能与实际环境不匹配。由此产生的模型失配可能导致求得的策略性能不佳。

这一挑战推动了针对鲁棒马尔可夫决策过程框架的开发与分析所开展的重要研究，例如（Bagnell、Ng 和 Schneider 2001；Nilim and El Ghaoui 2004；Iyengar 2005）。在鲁棒马尔可夫决策过程的设定中，人们并不采用固定的马尔可夫决策过程模型，而是寻求在可能的马尔可夫决策过程模型所构成的不确定集上优化最坏情况性能。该鲁棒马尔可夫决策过程问题的解能够为所有不确定的马尔可夫决策过程模型提供性能保证，因此能够抵御模型失配。

采用不同奖励最优性准则的鲁棒马尔可夫决策过程问题，在本质上有所不同。在鲁棒折扣马尔可夫决策过程中，目标是找到一个使最坏情况下的折扣累积奖励最大化的策略。在这种设置下，随着智能体与环境交互，获得的奖励会随时间呈指数级递减。以往鲁棒设置下的大量研究都聚焦于折扣奖励形式。基于模型的方法，例如（Iyengar 2005；Nilim 和 El Ghaoui 2004；Bagnell、Ng 和 Schneider 2001；Satia 和 Lave Jr 1973；Wiesemann、Kuhn 和 Rustem 2013；Tamar、Mannor 和 Xu 2014；Lim 和 Autef 2019；Xu 和 Mannor 2010；Yu 和 Xu 2015；Lim、Xu 和 Mannor，2013），假设学习者已知不确定集的信息，揭示了鲁棒折扣马尔可夫决策过程的若干基本特征。此后，研究进一步拓展到更具实践意义的无模型设定，在这种设定中，学习者只能获得来自模拟器（即不确定集的中心）的样本。例如，基于价值的方法（Roy、Xu 和 Pokutta 2017；Badrinath 和 Kalathil 2021；Wang 和 Zou 2021；Tessler、Efroni 和 Mannor 2019；Zhou 等，2021；Yang、Zhang 和 Zhang 2021；Panaganti 和 Kalathil 2021；Goyal 和 Grand-Clement 2018；Kaufman 和 Schaefer 2013；Ho、Petrik 和 Wiesemann 2018、2021；Si 等，2020）将鲁棒价值函数作为中间步骤来优化最坏情况性能；另一方面，无模型的基于策略方法（Russel、Benosman 和 Van Baar 2020；Derman、Geist 和 Mannor 2021；Eysenbach 和 Levine 2021；Wang 和 Zou 2022）直接优化策略，因此可扩展到大规模或连续的状态和动作空间。

尽管折扣马尔可夫决策过程具有优雅的贝尔曼算子，该算子是压缩映射，而且相关研究已经十分广泛，但当系统长时间运行时，由此获得的策略通常长期表现不佳。当折扣因子非常接近 1 时，智能体可能更倾向于根据平均期望奖励而非期望总折扣奖励来比较策略，例如排队控制、供应链中的库存管理、自动导引车调度以及通信网络中的应用（Kober、Bagnell 和 Peters，2013年）。因此，优化系统的长期平均

版权所有 © 2023，美国人工智能促进协会（www.aaai.org）。保留所有权利。

表现也很重要。

然而，平均奖励准则下的鲁棒马尔可夫决策过程鲜有研究。与折扣情形相比，平均奖励设定取决于底层随机过程的渐近行为，因此复杂得多。一个公认的复杂之处在于，平稳策略与状态—动作频率的极限点之间存在一一对应关系；这种关系在折扣马尔可夫决策过程中成立，但在平均奖励准则下，即便是在非鲁棒设定中，除少数非常特殊的情形外也不再成立（Puterman 1994；Atia 等，2021）。这主要是因为，在平均奖励设定中，建立压缩映射所需的条件取决于马尔可夫决策过程的图结构；而在折扣奖励设定中，只需折扣因子严格小于 1 即可。迄今为止，只有少数研究探讨了鲁棒设置下的平均奖励马尔可夫决策过程。Tewari 和 Bartlett（2007）的首项研究考察了特定有限区间不确定性集下的鲁棒平均奖励马尔可夫决策过程，但他们的方法不易推广至其他不确定性集。最近，Lim、Xu 和 Mannor（2013）针对 ℓ_1 不确定性集下的鲁棒平均奖励马尔可夫决策过程提出了一种算法。然而，要获得该问题的基本刻画并给出收敛保证，仍然困难重重。

挑战与贡献

本文针对具有一般不确定性集的鲁棒平均奖励马尔可夫决策过程，推导其表征，并提出具有可证明理论保证的基于模型的方法。我们的方法与以往关于鲁棒折扣马尔可夫决策过程、鲁棒和非鲁棒平均奖励马尔可夫决策过程的研究有本质区别。具体而言，主要挑战和贡献总结如下。

• **我们刻画了鲁棒折扣价值函数在折扣因子 $\gamma \to 1$下的渐近行为**。对于标准的非鲁棒设置和特定的转移核，折扣非鲁棒价值函数会收敛到平均奖励非鲁棒价值函数 $\gamma \to 1$ （Puterman 1994）。然而，在鲁棒设置下，我们需要考虑不确定性集内所有可能转移核下的最坏情况极限行为。因此，先前的逐点收敛结果（Puterman 1994）无法直接应用。在（Tewari 和 Bartlett 2007）中，研究了有限区间不确定性集；由于其特殊结构，鲁棒折扣马尔可夫决策过程的可能最坏情况转移核数量是有限的，因此可以交换 min（对转移核取最小值）和 $\lim_{\gamma \to 1}$ 的顺序，从而得到鲁棒折扣价值函数收敛到鲁棒平均奖励价值函数。然而，对于本文研究的一般不确定性集，该结果并不成立。我们首先证明，折扣非鲁棒价值函数关于转移核和策略一致收敛于平均奖励价值函数。在此一致收敛的基础上，我们证明鲁棒折扣价值函数收敛于鲁棒平均奖励价值函数。这一一致收敛结果是文献中的首次成果，对于推动我们的算法设计并保证其收敛到平均奖励设定下的最优鲁棒策略至关重要。

• **我们基于极限方法设计鲁棒策略评估和最优控制算法**。基于一致收敛，我们进而使用鲁棒折扣马尔可夫决策过程来近似鲁棒平均奖励马尔可夫决策过程。我们证明，当 γ 足够大时，鲁棒折扣马尔可夫决策过程的任一最优策略也是鲁棒平均奖励的最优策略，因此能够解决平均奖励设定下的鲁棒最优控制问题。这一结果类似于非鲁棒设定下的 Blackwell 最优性（Blackwell 1962年；Hordijk 和 Yushkevich 2002），但我们的证明在根本上有所不同。从技术上说，（Blackwell 1962年；Hordijk 和 Yushkevich 2002）中的证明基于这样一个事实：两个策略的折扣价值函数之差是折扣因子的有理函数，且其零点数量有限。然而，在具有一般不确定集的鲁棒设置下，由于对转移核取最小值，这一差值不再是有理函数。我们基于鲁棒折扣马尔可夫决策过程的渐近行为构造了一种新颖的证明，并证明当 $\gamma \to 1$时，（最优）鲁棒折扣价值函数收敛于（最优）鲁棒平均奖励。受这些发现启发，我们通过以一定速率逐步增大折扣因子，依次应用鲁棒折扣贝尔曼算子来设计算法。我们证明，该方法能够：（i）评估给定策略的鲁棒平均奖励；（ii）针对一般不确定性集，求得最优鲁棒价值函数，进而找到最优鲁棒策略。

• **我们设计了一种鲁棒相对价值迭代方法，无需将折扣马尔可夫决策过程作为中间步骤**。我们进一步采用直接方法，在不使用极限方法的情况下求解鲁棒平均奖励马尔可夫决策过程，即不将折扣马尔可夫决策过程作为中间步骤。我们为鲁棒平均奖励马尔可夫决策过程推导出鲁棒贝尔曼方程，并证明在平均奖励设定下，鲁棒相对价值函数与鲁棒平均奖励这一对量是鲁棒贝尔曼方程的一个解。我们进一步证明，如果能够找到鲁棒贝尔曼方程的任意一个解，就可以通过贪心方法推导出最优策略。因此，该问题等价于求解鲁棒贝尔曼方程。随后，我们设计了一种鲁棒值迭代方法，并证明该方法收敛于鲁棒贝尔曼方程的解，也就是求得鲁棒平均奖励马尔可夫决策过程问题的最优策略。

相关工作

鲁棒折扣马尔可夫决策过程。针对鲁棒折扣马尔可夫决策过程的基于模型的方法已在相关研究中得到探讨（Iyengar 2005；Nilim and El Ghaoui 2004；Bagnell、Ng 和 Schneider 2001；Satia 和 Lave Jr 1973；Wiesemann、Kuhn 和 Rustem 2013；Lim 和 Autef 2019；Xu 和 Mannor 2010；Yu 和 Xu 2015；Lim、Xu 和 Mannor，2013；Tamar、Mannor 和 Xu 2014）。这些研究假设不确定集已知，因此可以使用鲁棒动态规划求解问题。后来，相关研究被推广到无模型设定，在该设定中，可以通过在线方式获取来自不确定集的质心马尔可夫决策过程的随机样本（Roy、Xu 和 Pokutta 2017；Badrinath 和 Kalathil 2021；Wang 和 Zou 2021、2022；Tessler、Efroni 和 Mannor 2019），也可以通过离线方式获取（Zhou 等，2021；Yang、Zhang 和 Zhang 2021；Panaganti 和 Kalathil 2021；Goyal 和 Grand-Clement 2018；Kaufman 和 Schaefer 2013；Ho、Petrik 和 Wiesemann 2018、2021；Si 等，2020）。

此外，也有关于鲁棒强化学习的实证研究，例如（Vinitsky 等，2020；Pinto 等，2017；Abdullah 等，2019；Hou 等，2020；Rajeswaran 等，2017；Huang 等，2017；Kos 和 Song 2017；Lin 等，2017；Pattanaik 等，2018；Mandlekar 等，2017）。对于折扣型马尔可夫决策过程，鲁棒贝尔曼算子是一个压缩映射，基于此可以设计鲁棒动态规划和基于价值的方法。本文聚焦于鲁棒平均奖励马尔可夫决策过程。然而，平均奖励马尔可夫决策过程的鲁棒贝尔曼算子并非压缩映射，且其不动点可能不唯一。此外，平均奖励设定取决于底层随机过程的渐近行为，因此更加复杂。

鲁棒平均奖励马尔可夫决策过程。文献中关于鲁棒平均奖励马尔可夫决策过程的研究十分有限。Tewari 和 Bartlett（2007年）研究了具有特定有限区间不确定集的鲁棒平均奖励马尔可夫决策过程，并证明 Blackwell 最优策略存在，即存在某个 $\delta \in [0,1)$，使得最优鲁棒策略存在，且对任意折扣因子$\gamma \in [\delta, 1)$ 都保持不变。然而，这一结果依赖于不确定集的结构。对于一般不确定性集，Blackwell 最优策略的存在性可能无法得到保证。最近，Lim、Xu 和 Mannor（2013年）针对特定的 ℓ_1-范数不确定集设计了一种无模型算法，并刻画了其遗憾界。然而，他们的方法同样依赖于 ℓ_1-范数不确定集的结构，因此可能无法推广到其他类型的不确定性集。本文的结果适用于多种类型的不确定性集，因此具有更强的普适性。

预备知识与问题模型

本章节介绍折扣马尔可夫决策过程、平均奖励马尔可夫决策过程和鲁棒马尔可夫决策过程的一些预备知识。

折扣马尔可夫决策过程。一个折扣型马尔可夫决策过程 $(\mathcal{S}, \mathcal{A}, \mathrm{P}, r, \gamma)$ 由以下部分指定：状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、转移核 $\mathrm{P} = \{p_s^a \in \Delta(\mathcal{S})\ a \in \mathcal{A},\ s \in \mathcal{S}\}$[1]，其中 p_s^a 是采取动作 a 时在状态 s 下下一个状态在 $\mathcal{S}$ 上的分布（$p_{s,\,s'}^a$ 表示转移到 s' 的概率）；奖励函数 $r: \mathcal{S} \times \mathcal{A} \to [0,1]$；以及折扣因子 $\gamma \in [0,1)$。在每个时间步 t，处于状态 s_t的智能体采取动作 a_t，随后环境根据 $p_{s_t}^{a_t}$转移到下一个状态 s_{t+1}，并向智能体产生奖励信号 $r(s_t,\ a_t) \in [0,1]$。本文中，为方便起见，也将$r_t = r(s_t,\ a_t)$记作。

平稳策略 $\pi: \mathcal{S} \to \Delta(\mathcal{A})$ 是针对任意给定状态 s在$\mathcal{A}$上的一个分布，智能体在状态s 下以概率 $\pi(a|s)$ 执行动作 a。从 $s \in \mathcal{S}$ 开始的平稳策略 π 的折扣价值函数定义为遵循策略π：$V^{\pi}{}_{\mathrm{P}}{}_{\gamma}(s) \triangleq \mathbb{E}_{\pi,\ \mathrm{P}}[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s]$ 所获得的期望折扣累积奖励。

平均奖励马尔可夫决策过程。与折扣马尔可夫决策过程不同，平均奖励马尔可夫决策过程不会对随时间变化的回报进行折扣，而是考察底层马尔可夫过程在稳态分布下的行为。更具体地说，在给定转移核 P 的情况下，策略 π 从 $s \in \mathcal{S}$ 出发的平均回报定义为

$$g_{\mathrm{P}}^{\pi}(s) \triangleq \lim_{n\to\infty} \mathbb{E}_{\pi,\mathrm{P}}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right], \tag{1}$$

为方便起见，本文也将其称为平均奖励价值函数。

平均奖励价值函数也可以等价地写为：$g_{\mathrm{P}}^{\pi} = \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} (\mathrm{P}^{\pi})^t r_{\pi} \triangleq \mathrm{P}_*^{\pi} r_{\pi}$，其中$(\mathrm{P}^{\pi})_{s,\,s'} \triangleq \sum_a \pi(a|s) p_{s,\,s'}^a$ 以及 $r_{\pi}(s) \triangleq \sum_a \pi(a|s) r(s,\ a)$ 分别是由 π 诱导的转移矩阵和奖励函数，而 $\mathrm{P}_*^{\pi} \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} (\mathrm{P}^{\pi})^t$ 是 P^{π} 的极限矩阵。

在平均奖励设定中，我们还定义如下相对价值函数

$$V_{\mathrm{P}}^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathrm{P}}\left[\sum_{t=0}^{\infty} (r_t - g_{\mathrm{P}}^{\pi}) | S_0 = s\right], \tag{2}$$

它表示回报与平均值 g_{P}^{π} 之间随时间累积的差值。已有研究证明（Puterman 1994）：$V^{\pi}{}_{\mathrm{P}} = H_{\mathrm{P}}^{\pi} r_{\pi}$，其中 $H_{\mathrm{P}}^{\pi} \triangleq (I - \mathrm{P}^{\pi} + \mathrm{P}_*^{\pi})^{-1}(I - \mathrm{P}_*^{\pi})$ 定义为 P^{π} 的偏差矩阵。

平均回报函数与相对价值函数之间的关系可以用以下贝尔曼方程来描述（Puterman 1994）：

$$V_{\mathrm{P}}^{\pi}(s) = \mathbb{E}_{\pi}\left[r(s,A) - g_{\mathrm{P}}^{\pi}(s) + \sum_{s' \in \mathcal{S}} p_{s,s'}^{A} V_{\mathrm{P}}^{\pi}(s')\right]. \tag{3}$$

鲁棒折扣型与平均回报型马尔可夫决策过程。对于鲁棒马尔可夫决策过程，转移核并非固定，而是属于某个不确定集 $\mathcal{P}$。智能体执行动作后，环境会根据任意转移核 $\mathrm{P} \in \mathcal{P}$ 转移到下一个状态。本文聚焦于 $(s,\ a)$- 矩形不确定集（Nilim and El Ghaoui 2004；Iyengar 2005），即 $\mathcal{P} = \bigotimes_{s,\ a} \mathcal{P}_s^a$，其中 $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$。我们注意到，也有研究探讨将 $(s,\ a)$-矩形不确定集放宽为 s-矩形不确定集，但这并非本文的研究重点。

在鲁棒设置下，我们考虑马尔可夫决策过程不确定集上的最坏情况表现。更具体地说，折扣马尔可夫决策过程的策略 π 的鲁棒折扣价值函数定义为

$$V_{\mathcal{P},\gamma}^{\pi}(s) \triangleq \min_{\kappa \in \bigotimes_{t\geq 0} \mathcal{P}} \mathbb{E}_{\pi,\kappa}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right], \tag{4}$$

其中 $\kappa = (\mathrm{P}_0, \mathrm{P}_1 \ldots) \in \bigotimes_{t\geq 0} \mathcal{P}$。

[1] $\Delta(\mathcal{S})$：$\mathcal{S}$上的（$|\mathcal{S}| - 1$）维概率单纯形。

在本文中，我们关注策略 π 的以下最坏情况平均奖励:

$$g_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\pi,\kappa}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right], \quad (5)$$

为方便起见，我们将其称为鲁棒平均奖励价值函数。

对于鲁棒折扣马尔可夫决策过程，已有研究证明，鲁棒折扣价值函数是鲁棒折扣贝尔曼算子的唯一不动点（Nilim and El Ghaoui 2004；Iyengar 2005；Puterman 1994）:

$$\mathbf{T}_{\pi}V(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s)\left(r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V)\right), \quad (6)$$

其中，$\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{p\in\mathcal{P}_s^a} p^{\top}V$是 $\mathcal{P}_s^a$上V的支撑函数。基于$\mathbf{T}_{\pi}$的压缩映射性质，可以设计鲁棒动态规划方法，例如鲁棒值迭代（Nilim and El Ghaoui 2004；Iyengar 2005）（有关这些方法的回顾，请参阅附录）。然而，鲁棒平均奖励马尔可夫决策过程并不存在这样的压缩映射结果。本文旨在寻找一个能够优化鲁棒平均奖励价值函数的策略:

$$\max_{\pi\in\Pi} g_{\mathcal{P}}^{\pi}(s), \text{ for any } s \in \mathcal{S}, \quad (7)$$

其中，Π 是所有平稳策略的集合，我们将 $g_{\mathcal{P}}^{*}(s) \triangleq \max_{\pi} g_{\mathcal{P}}^{\pi}(s)$记为最优鲁棒平均奖励。

鲁棒平均奖励马尔可夫决策过程的极限方法

我们首先采用极限方法来解决式（7）中的鲁棒平均奖励马尔可夫决策过程问题。众所周知，在非鲁棒设定下，对于任意固定的π和P，当折扣因子γ趋近于1时，折扣价值函数会收敛到平均奖励价值函数（Puterman 1994），即，

$$\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^{\pi} = g_{\mathsf{P}}^{\pi}. \quad (8)$$

我们采用类似的思路，并证明在鲁棒情形下同样成立：在一个温和的假设下，$\lim_{\gamma\to 1}(1-\gamma)V_{\mathcal{P},\gamma}^{\pi} = g_{\mathcal{P}}^{\pi}$。基于这一结果，我们进一步设计了算法（算法1和算法2），以一定的速率逐渐增大折扣因子，同时依次应用鲁棒折扣贝尔曼算子。随后，我们从理论上证明了这些算法会收敛到最优解。

接下来，我们首先证明收敛性 $\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^{\pi} = g_{\mathsf{P}}^{\pi}$ 在集合 $\Pi \times \mathcal{P}$ 上是一致的。在平均奖励马尔可夫决策过程的研究中，通常会考虑某些类别的马尔可夫决策过程，例如单链和连通马尔可夫决策过程（Wei等人 2020年；Zhang和Ross 2021；Chen、Jain和Luo 2022年；Wan、Naik和Sutton 2021）。本文聚焦于单链设定，以突出实现鲁棒性的主要技术创新。

假设1。 对于任意 $s \in \mathcal{S}$、$a \in \mathcal{A}$，不确定集$\mathcal{P}_s^a$是 $\Delta(\mathcal{S})$ 的紧致子集。并且对于任意 $\pi \in \Pi$，$\mathsf{P} \in \mathcal{P}$，诱导*MDP*是单链的。

假设1的第一部分相当于假设不确定集是闭集。我们注意到，许多标准的不确定集都满足这一假设，例如由 ϵ污染（Huber 1965）、有限区间（Tewari and Bartlett 2007）、全变差（Rahimian, Bayraksan, and De-Mello 2022）和KL散度（Hu and Hong 2013）定义的不确定集。单链假设也广泛用于平均奖励马尔可夫决策过程的研究，例如（Puterman 1994；Wan、Naik和Sutton 2021；Zhang和Ross 2021；Lan 2020；Zhang, Zhang, and Maguluri 2021）。此外，值得注意的是，在单链假设下，对于每个初始状态，鲁棒平均奖励都相同，即 $g_{\mathsf{P}}^{\pi}(s_1) = g_{\mathsf{P}}^{\pi}(s_2)\ \forall s_1,\ s_2 \in \mathcal{S}$（Bertsekas 2011）。

注 1。 本节的结果实际上只要求 $\|H_{\mathsf{P}}^{\pi}\|$、$\forall\pi \in \Pi$ 和 $\mathsf{P} \in \mathcal{P}$ 一致有界（见附录中的引理 *2*）。假设 *1* 是一个充分条件。

在（Puterman 1994）中，对于固定策略 π 和固定转移核 P（非鲁棒设定），收敛 $\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^{\pi} = g_{\mathsf{P}}^{\pi}$ 是逐点收敛。然而，这种逐点收敛并不能保证鲁棒折扣价值函数收敛，因为鲁棒价值函数衡量的是不确定集上的最坏情况性能，而一般而言，lim 和 min 的运算顺序不能互换。在下面的定理中，我们证明，在上述假设下，折扣价值函数一致收敛。

定理 1 （一致收敛）。 在假设*1*下，当 $\gamma \to 1$ 时，折扣价值函数在 $\Pi \times \mathcal{P}$ 上一致收敛于平均奖励价值函数，即

$$\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^{\pi} = g_{\mathsf{P}}^{\pi}, \textit{ uniformly.} \quad (9)$$

根据定理 1 的一致收敛性，极限 $\gamma \to 1$ 和 $\min_{\mathsf{P}}$ 的顺序可以互换，进而可以证明鲁棒折扣价值函数具有如下收敛性。

定理 2. 式 (4)中的鲁棒折扣价值函数在 Π上一致收敛于鲁棒平均奖励:

$$\lim_{\gamma\to 1}(1-\gamma)V_{\mathcal{P},\gamma}^{\pi} = g_{\mathcal{P}}^{\pi} \textit{ uniformly.} \quad (10)$$

我们注意到，（Tewari and Bartlett 2007）中证明了类似的收敛结果，但仅适用于有限区间这一特殊不确定集。我们的定理 2 适用于一般紧不确定性集。此外，值得强调的是，我们的证明方法与（Tewari and Bartlett 2007）中的方法有本质区别。具体而言，在有限区间不确定性集下，最坏情况转移核来自一个有限集，即 $V_{\mathcal{P},\gamma}^{\pi} = \min_{\mathsf{P}\in\mathcal{M}} V^{\pi}{}_{\mathsf{P}\,\gamma}$，其中$\mathcal{M} \subseteq \mathcal{P}$为有限集。因此，lim 和 min 可以互换。然而，对于一般不确定性集，最坏情况转移核的数量可能不是有限的。我们通过定理 1 的一致收敛结果证明了二者可以互换。

前述两个收敛结果在鲁棒平均奖励马尔可夫决策过程的极限方法中发挥着基础性作用，对于激发以下

两种算法的设计至关重要。其基本思想是从任意初始值出发，按一定速率逐步增大折扣因子，并依次应用一系列鲁棒折扣贝尔曼算子。

我们首先考虑鲁棒策略评估问题，其目标是估计固定策略 π 下的鲁棒平均回报 $g_{\mathcal{P}}^{\pi}$。关于鲁棒折扣马尔可夫决策过程的这一问题，已有大量文献研究；然而，关于鲁棒平均奖励马尔可夫决策过程的研究结果十分有限，只有 Tewari and Bartlett 2007 针对特定有限区间不确定性集的研究例外。我们在算法 1 中提出用于一般不确定性集的鲁棒值迭代（鲁棒 VI）算法，以评估鲁棒平均回报。

算法 1：鲁棒 VI：策略评估

输入：π，$V_0(s)=0$，$\forall s$，T

对于 $t=0,1,\ldots,T-1$ **执行**

  $\gamma_t \leftarrow \frac{t+1}{t+2}$

  对于所有 $s\in\mathcal{S}$，**执行**

    $V_{t+1}(s) \leftarrow \mathbb{E}_{\pi}[(1-\gamma_t)\, r(s,A) + \gamma_t \sigma_{\mathcal{P}^{A}_{s}}(V_t)]$

  结束循环

结束循环

返回 V_T

在每个时间步 t，折扣因子 γ_t 设为 $\frac{t+1}{t+2}$，当 $t\to\infty$时其收敛到 1。随后，将关于折扣因子 γ_t 的鲁棒贝尔曼算子应用于鲁棒折扣价值函数 $(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}$ 的当前估计 V_t。随着折扣因子趋近于 1，根据定理 2，估计的鲁棒折扣价值函数收敛到鲁棒平均奖励 $g_{\mathcal{P}}^{\pi}$。以下结果表明，算法 1 的输出收敛到鲁棒平均奖励。

定理 3。 算法 *1* 收敛到鲁棒平均奖励，即 $\lim_{T\to\infty} V_T = g_{\mathcal{P}}^{\pi}$。

除了鲁棒策略评估问题之外，找到一个使最坏情况平均奖励最大化的最优策略（即求解式（7））在实践中也具有重要意义。基于与算法 1 类似的思路，我们将极限方法扩展到算法 2，以求解鲁棒最优控制问题。

算法 2：鲁棒值迭代：最优控制

输入：$V_0(s)=0$，$\forall s$，T

对于 $t=0,1,\ldots,T-1$ **执行**

  $\gamma_t \leftarrow \frac{t+1}{t+2}$

  对于所有 $s\in\mathcal{S}$**执行**

    $V_{t+1}(s) \leftarrow \max_{a\in\mathcal{A}} \left\{(1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^{a}_{s}}(V_t)\right\}$

  结束循环

结束循环

对于 $s\in\mathcal{S}$**执行**

  $\pi_T(s) \leftarrow$ 最大值参数$_{a\in\mathcal{A}} \left\{(1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^{a}_{s}}(V_T)\right\}$

结束循环

返回 V_T, π_T

与算法 1 类似，在每个时间步，折扣因子 γ_t 会逐渐趋近于 1，并将相对于 γ_t 的单步鲁棒折扣贝尔曼算子（用于最优控制）应用于当前估计 V_t。以下定理证明，算法 2 中的 V_T 收敛到最优鲁棒价值函数，因此能够找到最优鲁棒策略。

定理 4. 算法 *2* 的输出 V_T 收敛到最优鲁棒平均奖励 $g_{\mathcal{P}}^{*}$：当 $T\to\infty$时，$V_T \to g_{\mathcal{P}}^{*}$。

正如 (Blackwell 1962; Hordijk 和 Yushkevich 2002) 中所讨论的，平均奖励准则并不敏感且具有选择性，因为它只关注稳态分布下的表现。例如，获得奖励 $100+0+0+\cdots$ 和 $0+0+0+\cdots$的两个策略同样好/差。针对这一问题，在非鲁棒设定中，Blackwell 引入了一种更敏感的最优性概念 (Blackwell 1962)。具体而言，如果某个策略能够针对某个 $\delta\in$ (0,1) 所对应的所有折扣因子 $\gamma\in(\delta,1)$ 优化折扣价值函数，那么该策略就称为 Blackwell 最优策略。结合公式 (8)，通过取 $\gamma\to 1$ 得到的最优策略不仅对于平均奖励准则是最优的，对于较大的 γ所对应的折扣准则也是最优的。直观来说，它在平均奖励设定下是最优的，同时也重视早期奖励。

遵循类似的思路，我们证明，算法 2 得到的策略不仅在鲁棒平均回报设定下是最优的，而且对早期奖励也很敏感。

用 Π_D^{*} 表示所有鲁棒平均奖励下的确定性最优策略的集合（引理 7 已证明该集合非空），即 $\Pi_D^{*}=\{\pi\in\Pi_D:\, g_{\mathcal{P}}^{\pi}=g_{\mathcal{P}}^{*}\}$.

定理 5 (Blackwell最优性)。 存在 $0<\delta<1$，使得对于任意 $\gamma>\delta$，鲁棒折扣价值函数 $V^{*}_{\mathcal{P},\gamma}$ 对应的确定性最优鲁棒策略都属于Π_D^{*}。此外，当 Π_D^{*} 为单元素集合时，存在唯一的 *Blackwell* 最优策略。

这一结果表明，使用本节中的极限方法寻找平均奖励马尔可夫决策过程的最优鲁棒策略还有一个额外优势：该方法找到的策略不仅能使稳态下的平均奖励达到最优，而且对早期奖励也很敏感。

值得强调的是，我们的结果与证明 Blackwell 最优性（Blackwell 1962）时所用技巧之间的区别。在非鲁棒设定中，平稳 Blackwell 最优策略的存在性是通过矛盾法证明的，证明中用到了两个策略 π 和 ν:$f_{\pi,\nu}(\gamma)\triangleq V^{\pi}_{\mathcal{P},\gamma}-V^{\mu}_{\mathcal{P},\gamma}$的差函数。通过矛盾法证明了 f 有无穷多个零点，但这与 f 是关于 γ的有理函数且零点个数有限这一事实相矛盾。Tewari 和 Bartlett （2007）在有限区间不确定性集的情形下也使用了类似技巧。具体而言，Tewari 和 Bartlett（2007）证明，对于任意 π，γ，最坏情况转移核均来自一个有限集合 $\mathcal{M}$，因此 $f_{\pi,\nu}(\gamma)\triangleq \min_{\mathcal{P}\in\mathcal{M}} V^{\pi}_{\mathcal{P},\gamma} - \min_{\mathcal{P}\in\mathcal{M}} V^{\mu}_{\mathcal{P},\gamma}$也可以证明是有理函数，且零点个数有限。然而，对于一般的不确定集 $\mathcal{P}$，差函数 $f_{\pi,\nu}(\gamma)$ 可能不是有理函数。这使得（Blackwell 1962；Tewari and Bartlett 2007）中的方法不适用于我们的问题。

鲁棒平均奖励马尔可夫决策过程的直接方法

第 章节中的极限方法基于折扣价值函数的一致收敛，并使用折扣马尔可夫决策过程来近似平均奖励马尔可夫决策过程。本节将介绍一种直接求解鲁棒平均奖励马尔可夫决策过程的方法，该方法不采用折扣马尔可夫决策过程作为中间步骤。

对于平均奖励马尔可夫决策过程，通常采用相对价值迭代（Puterman 1994）方法，因为该方法具有数值稳定性和收敛保证。下面，我们将相对价值迭代算法推广到鲁棒设置，并在算法3中设计鲁棒相对价值迭代算法。

我们首先将式（2）中的相对价值函数推广为鲁棒相对价值函数。对于策略 π，鲁棒相对价值函数衡量最坏情况累积奖励与最坏情况平均奖励之间的差异。

定义1。 鲁棒相对价值函数定义为

$$V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa \in \bigotimes_{t\geq 0} \mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right], \quad (11)$$

其中， $g_{\mathcal{P}}^{\pi}$ 是式中定义的最坏情况平均奖励。 (5).

以下定理给出了鲁棒平均奖励马尔可夫决策过程的鲁棒贝尔曼方程。

定理 6. 对于任意 s 和 π，$(V_{\mathcal{P}}^{\pi}, g_{\mathcal{P}}^{\pi})$ 都是下列鲁棒贝尔曼方程的解：

$$V(s) + g = \sum_a \pi(a|s)\left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V)\right). \quad (12)$$

可以看出，平均回报马尔可夫决策过程的鲁棒贝尔曼方程与式 (6) 中折扣型马尔可夫决策过程的方程结构相似，只是前者没有折扣因子。这实际上揭示了折扣型马尔可夫决策过程与平均回报马尔可夫决策过程的鲁棒贝尔曼算子之间的一个根本区别。对于折扣型马尔可夫决策过程，其鲁棒贝尔曼算子是压缩映射，压缩常数为 γ （Nilim and El Ghaoui 2004；Iyengar 2005），因此不动点是唯一的。基于这一点，可以通过递归应用鲁棒贝尔曼算子来求得鲁棒价值函数（见附录）。与此形成鲜明对比的是，在平均奖励设定下，鲁棒贝尔曼算子不一定是压缩映射，而且不动点可能不唯一。因此，在平均奖励设定下反复应用鲁棒贝尔曼算子甚至可能无法收敛，这凸显了这两种问题设定在本质上的不同。

我们首先推导出鲁棒平均奖励马尔可夫决策过程的以下等价最优性条件。

定理7。对于任意满足条件的(g, V) *to*

$$\max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) - V(s)\right\} = 0, \forall s, \quad (13)$$

$g = g_{\mathcal{P}}^{*}$.如果我们进一步设定

$$\pi^*(s) = \arg\max_a \left\{r(s,a) + \sigma_{\mathcal{P}_s^a}(V)\right\} \quad (14)$$

对于任意 $s \in \mathcal{S}$，那么 π^* 是一个最优鲁棒策略。

定理7表明，只要我们找到方程（13）的一个解（g, V），尽管该解可能并不唯一，那么 g 就是最优鲁棒平均奖励 $g_{\mathcal{P}}^{*}$，而贪心策略π^* 则是我们在方程（7）中的鲁棒平均回报马尔可夫决策过程问题的最优策略。

下文中，我们将相对值迭代方法推广到鲁棒设置，并在算法3中设计一种鲁棒相对值迭代算法。我们还将证明，该算法的输出收敛到方程（13）的一个解，并且可以通过方程（14）得到最优策略。这里 $\mathbb{1}$ 定——

算法3：鲁棒 RVI

输入：V_0、ϵ 以及任意的 $s^* \in \mathcal{S}$

1: $w_0 \leftarrow V_0 - V_0(s^*)\mathbb{1}$
2: **当** $sp(w_t - w_{t+1}) \geq \epsilon$ **时**
3: **对于**所有 $s \in \mathcal{S}$
4: $V_{t+1}(s) \leftarrow$ 最大$_a(r(s,a) + \sigma_{\mathcal{P}_s^a}(w_t))$
5: $w_{t+1}(s) \leftarrow V_{t+1}(s) - V_{t+1}(s^*)$
6: **结束循环**
7: **结束 while 循环**
8: **返回** w_t, V_t

表示全一向量，而 sp 表示跨度半范数：$sp(w) = \max_s w(s) - \min_s w(s)$。与算法 2 不同，在算法3中，我们不需要应用鲁棒折扣贝尔曼算子。该方法直接求解平均奖励鲁棒马尔可夫决策过程的鲁棒最优控制问题。

为了研究鲁棒相对值迭代算法的收敛性，我们首先作出如下附加假设。

假设 2. 存在一个正整数 J，使得对于任意 $\mathsf{P} = \{p_s^a \in \Delta(\mathcal{S})\} \in \mathcal{P}$ 以及任意平稳确定性策略 π，都存在 $\kappa > 0$ 和一个状态 $s \in \mathcal{S}$，使得$((\mathsf{P}^{\pi})^J)_{x,s} \geq \kappa, \forall x \in \mathcal{S}$。

这个假设已被证明等价于假设系统是单链且非周期的（Bertsekas 2011）。也可以用一些较弱的假设替代它，例如（Bertsekas 2011）的命题 4.3.2；或者通过设计相对值迭代算法的一个变体将其移除，例如（Bertsekas 2011）的命题 4.3.4。在下面的定理中，我们证明算法3收敛到方程（13）的一个解；因此，根据定理7，若按照（14）设置 π，则 π 是最优鲁棒策略。

定理 8. (w_b, V_t) 收敛到方程(w, V) 的一个解(13)，当 $\epsilon \to 0$ 时。

备注 2. 本节主要介绍用于鲁棒最优控制问题的鲁棒相对值迭代算法，以及其收敛性和最优性保证。将算法*3*第*4*行中的$\max$ 替换为关于 π的期望，即可类似地设计用于鲁棒策略评估的鲁棒相对值迭代算法。定理 *8* 中的收敛结果也可类似地推导。

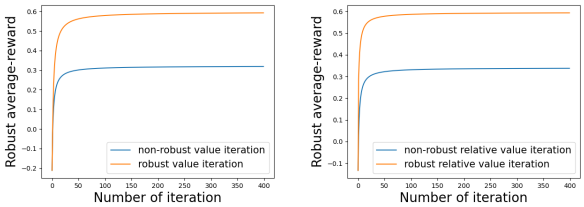
示例与数值结果

在本章节中，我们研究几种常用的不确定集模型，包括污染模型、Kullback-Lerbler（KL）散度和总变差定义模型。

从算法1至3可以看出，对于不同的不确定性集，唯一的区别在于支撑函数 $\sigma_{\mathcal{P}_s^a}(V)$ 的计算方式。接下来，我们将讨论如何高效计算各种不确定性集的支撑函数。

我们对不同不确定集下的鲁棒（相对）值迭代方法与非鲁棒（相对）值迭代方法进行了数值比较。实验基于 Garnet 问题 $\mathcal{G}(20, 40)$（Archibald、McKinnon 和 Thomas，1995）。具体而言，问题包含 20 个状态和 30 个动作；名义转移核 $\mathrm{P} = \{p_s^a \in \Delta(\mathcal{S})\}$根据均匀分布随机生成，奖励函数为 $r(s, a) \sim \mathcal{N}(0, \sigma_{s,a})$，其中$\sigma_{s,a} \sim$ 均匀分布$[0, 1]$。在实验中，不确定集以名义转移核为中心构造。我们运行不同的算法，即（鲁棒）值迭代和（鲁棒）相对值迭代，并在每个时间步获取贪心策略。随后，使用鲁棒平均回报策略评估（算法 1）评估这些策略的鲁棒平均回报。我们绘制鲁棒平均回报随迭代次数变化的曲线。

污染模型。 对于任意 (s, a)，不确定集 $\mathcal{P}_s^a$定义为 $\mathcal{P}_s^a = \{q : q = (1-R)p_s^a + Rp', p' \in \Delta(\mathcal{S})\}$，其中 p_s^a 是名义转移核。它可以看作一种对抗模型：在每个时间步，环境以 $1-R$的概率按照名义转移核 p 转移，以 R 的概率按照任意转移核 p' 转移。注意，$\sigma_{\mathcal{P}_s^a}(V) = (1-R)(p_s^a)^\top V + R\min_s V(s)$。 污染模型下的实验结果如图1所示。



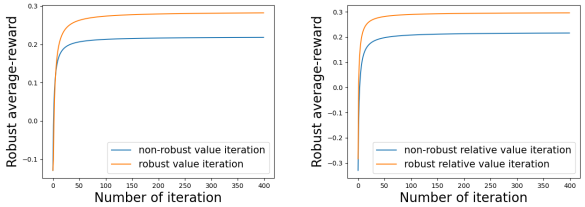
(a) 鲁棒值迭代。 (b) 鲁棒 RVI。

图1： $R = 0.4$条件下污染模型的比较。

全变差。 全变差距离是另一种常用的距离度量，用于衡量两个分布之间的差异。对于两个分布 p和 q，其定义为 $D_{TV}(p, q) = \frac{1}{2}\|p - q\|_1$.考虑通过全变差定义的不确定集： $\mathcal{P}_s^a = \{q : D_{TV}(q\|p_s^a) \le R\}$。然后，其支撑函数可以通过以下方式高效求解（Iyengar 2005 $\sigma_{\mathcal{P}_s^a} V = p^\top V -$）：$R$ V s s $-$ V s s ( ) $\min_{\mu \ge 0}\{\max_s(\ ) - \mu(\ )) \min_s(\ ) - \mu(\ ))\}$.

我们在全变差模型下的实验结果如图2所示。

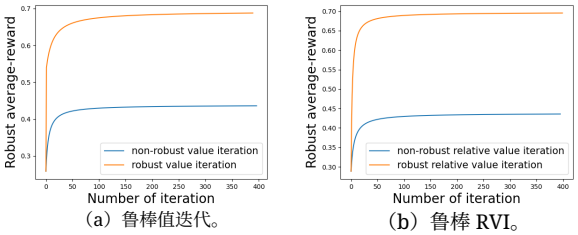
Kullback-Lerbler（KL）散度。Kullback–Leibler 散度广泛用于衡量两个概率分布之间的距离。对于分布 p, q，其定义为 $D_{KL}(q\|p) = \sum_s q(s)\log\frac{q(s)}{p(s)}$. 考虑通过 KL 散度定义的不确定集： $\mathcal{P}_s^a = \{q : D_{KL}(q\|p_s^a) \le R\}$。然后，可以利用（Hu and Hong



（a）鲁棒值迭代。 （b）鲁棒 RVI。

图2： $R = 0.6$下全变差模型的比较。

2013年）：$\sigma_{\mathcal{P}_s^a}(V) = -\min_{\alpha \ge 0}\left\{R\alpha + \alpha\log\left(p^\top e^{\frac{-V}{\alpha}}\right)\right\}$.我们在 KL 散度模型下的实验结果如图 3 所示。



（a）鲁棒值迭代。 （b）鲁棒 RVI。

Figu图 3：KL散度模型上的比较 $R = 0.8$。

可以看出，我们的鲁棒方法能够获得最坏情况奖励更高的策略。此外，我们基于极限方法的鲁棒值迭代和鲁棒相对价值迭代的直接方法都收敛到最优鲁棒策略，这验证了我们的理论结果。

结论

本文研究了平均奖励设定下的鲁棒马尔可夫决策过程问题。我们证明了折扣价值函数一致收敛到平均奖励值，这进一步意味着鲁棒折扣价值函数一致收敛到鲁棒平均奖励值。基于这一发现，我们设计了一种鲁棒动态规划方法，以鲁棒折扣马尔可夫决策过程作为近似（极限方法）。我们从理论上证明了该方法的收敛性和最优性，并证明了鲁棒版本的 Blackwell 最优性（Blackwell 1962）。随后，我们为鲁棒平均奖励马尔可夫决策过程设计了一种直接方法，并推导出其鲁棒贝尔曼方程。我们还设计了一种鲁棒 RVI 方法，并证明其收敛到最优鲁棒解。在技术层面上，我们的证明方法与现有关于平均奖励鲁棒马尔可夫决策过程的研究有根本区别，例如（Blackwell 1962；Tewari and Bartlett 2007）中的研究。

致谢

本研究得到美国国家科学基金会资助，项目编号为 CCF-2106560、CCF-2007783、CCF-2106339 和 CCF-1552497。

参考文献

Abdullah, M. A.; Ren, H.; Ammar, H. B.; Milenkovic, V.; Luo, R.; Zhang, M.; and Wang, J. 2019. Wasserstein鲁棒强化学习。*arXiv*预印本 *arXiv:1907.13196*。Archibald, T.; McKinnon, K.; and Thomas, L. 1995. 关于马尔可夫决策过程的生成。《运筹学会杂志》，46(3)：354–361。Atia, G. K.; Beckus, A.; Alkhouri, I.; and Velasquez, A. 2021. 期望回报多链马尔可夫决策过程的稳态规划。《人工智能研究杂志》，72：1029–1082。Badrinath, K. P.; and Kalathil. 2021. 使用最小二乘策略迭代并提供可证明性能保证的鲁棒强化学习。载于《国际机器学习会议（*ICML*）论文集》，511–520。PMLR。Bagnell, J. A.; Ng, A. Y.; and Schneider, J. G. 2001. 求解不确定马尔可夫决策过程。Bertsekas, D. P. 2011.《动态规划与最优控制》第3版，第II卷。马萨诸塞州贝尔蒙特：*Athena Scientific*。Blackwell, D. 1962. 离散动态规划。《数理统计年刊》，719–726。Chen, L.; Jain, R.; and Luo, H. 2022. 带约束的无限时域平均奖励马尔可夫决策过程学习。*arXiv*预印本 *arXiv:2202.00150*。Derman, C. 1970. 有限状态马尔可夫决策过程。Academic Press, Inc. Derman, E.; Geist, M.; and Mannor, S. 2021. 双重正则化马尔可夫决策过程，以及鲁棒性与正则化之间的等价性。载于《神经信息处理系统进展》（*NeurIPS*）论文集。Eysenbach, B.; and Levine, S. 2021. 最大熵强化学习（可证明地）能够解决某些鲁棒强化学习问题。*arXiv*预印本 *arXiv:2103.06257*。Goyal, V.; and Grand-Clement, J. 2018. 鲁棒马尔可夫决策过程：超越矩形性。*arXiv*预印本 *arXiv:1811.00215*。Ho, C. P.; Petrik, M.; and Wiesemann, W. 2018. 鲁棒马尔可夫决策过程的快速贝尔曼更新。载于《国际机器学习会议（*ICML*）论文集，1979–1988。PMLR。Ho, C. P.; Petrik, M.; and Wiesemann, W. 2021. L1鲁棒马尔可夫决策过程的部分策略迭代。《机器学习研究杂志》，22(275)：1–46。Hordijk, A.; and Yushkevich, A. A. 2002. Blackwell最优性。载于《马尔可夫决策过程手册》，231–267。Springer。Hou, L.; Pang, L.; Hong, X.; Lan, Y.; Ma, Z.; and Yin, D. 2020. 具有Wasserstein约束的鲁棒强化学习。*arXiv*预印本 *arXiv:2006.00945*。Hu, Z.; and Hong, L. J. 2013. 受Kullback-Leibler散度约束的分布鲁棒优化。可从*Optimization Online*获取，1695–1724。Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; and Abbeel, P. 2017. 针对神经网络策略的对抗攻击。载于《国际学习表征会议（*ICLR*）论文集。

Huber, P. J. 1965年。《概率比检验》的鲁棒版本。《数学统计年刊》，36：1753–1758。Iyengar, G. N. 2005年。鲁棒动态规划。《运筹学数学》，30(2)：257–280。Kaufman, D. L.；和 Schaefer, A. J. 2013年。鲁棒修正策略迭代。《*INFORMS*计算杂志》，25(3)：396–410。Kober, J.；Bagnell, J. A.；和 Peters, J. 2013年。机器人学中的强化学习：综述。《国际机器人研究杂志》，32(11)：1238–1274。Kos, J.；和 Song, D. 2017年。深入探究针对深度策略的对抗攻击。载于国际学习表征会议（*ICLR*）论文集。Lan, G. 2020年。机器学习的一阶与随机优化方法。施普林格·自然。Lim, S. H.；和 Autef, A. 2019年。鲁棒马尔可夫决策过程中基于核的强化学习。载于国际机器学习会议（*ICML*）论文集，3973–3981。PMLR。Lim, S. H.；Xu, H.；和 Mannor, S. 2013年。鲁棒马尔可夫决策过程中的强化学习。载于神经信息处理系统进展（*NIPS*）论文集，701–709。Lin, Y.-C.；Hong, Z.-W.；Liao, Y.-H.；Shih, M.-L.；Liu, M.-Y.；和 Sun, M. 2017年。针对深度强化学习智能体的对抗攻击策略。载于国际人工智能联合会议（*IJCAI*）论文集，3756–3762。Mandlekar, A.；Zhu, Y.；Garg, A.；Fei-Fei, L.；和 Savarese, S. 2017年。对抗鲁棒策略学习：物理上合理扰动的主动构造。载于*2017*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），3932–3939。IEEE。Nilim, A.；和 El Ghaoui, L. 2004年。具有不确定转移矩阵的马尔可夫决策问题中的鲁棒性。载于神经信息处理系统进展（*NIPS*）论文集，839–846。Panaganti, K.；和 Kalathil, D. 2021年。具有生成模型的鲁棒强化学习的样本复杂度。*arXiv*预印本 *arXiv:2112.01506*。Pattanaik, A.；Tang, Z.；Liu, S.；Bommannan, G.；和 Chowdhary, G. 2018年。对抗攻击下的鲁棒深度强化学习。载于自主智能体与多智能体系统国际会议论文集，2040–2042。Pinto, L.；Davidson, J.；Sukthankar, R.；和 Gupta, A. 2017年。鲁棒对抗强化学习。载于国际机器学习会议（*ICML*）论文集，2817–2826。PMLR。Puterman, M. L. 1994年。《马尔可夫决策过程：离散随机动态规划》。Rahimian, H.；Bayraksan, G.；和 De-Mello, T. H. 2022年。关注全变差距离的多阶段分布鲁棒优化中的有效情景。《*SIAM*优化杂志》，32(3)：1698–1727。Rajeswaran, A.；Ghotra, S.；Ravindran, B.；和 Levine, S. 2017年。Epopt：利用模型集成学习鲁棒神经网络策略

。载于国际学习表征会议（*ICLR*）论文集。Roy, A.；Xu, H.；和 Pokutta, S. 2017年。模型失配下的强化学习。载于神经信息处理系统进展（*NIPS*）论文集，3046–3055。Rudin, W. 2022年。泛函分析。McGraw-Hill Science &Engineering &Math，第2版。Russel, R. H.；Benosman, M.；和 Van Baar, J. 2020年。鲁棒约束马尔可夫决策过程：模型不确定性下的软约束鲁棒策略优化。*arXiv*预印本 *arXiv:2010.04870*。Satia, J. K.；和 Lave Jr, R. E. 1973年。具有不确定转移概率的马尔可夫决策过程。运筹学，21(3)：728–740。Si, N.；Zhang, F.；Zhou, Z.；和 Blanchet, J. 2020年。离线上下文赌博机中的分布鲁棒策略评估与学习。载于国际机器学习会议（*ICML*）论文集，8884–8894。PMLR。Sigaud, O.；和 Buffet, O. 2013年。人工智能中的马尔可夫决策过程。约翰·威立父子出版公司。Sutton, R. S.；和 Barto, A. G. 2018年。强化学习：导论。马萨诸塞州剑桥：麻省理工学院出版社。Tamar, A.；Mannor, S.；和 Xu, H. 2014年。利用函数逼近扩展鲁棒马尔可夫决策过程。载于国际机器学习会议（*ICML*）论文集，181–189。PMLR。Tessler, C.；Efroni, Y.；和 Mannor, S. 2019年。动作鲁棒强化学习及其在连续控制中的应用。载于国际机器学习会议，6215–6224。PMLR。Tewari, A.；和 Bartlett, P. L. 2007年。具有平均奖励准则的有界参数马尔可夫决策过程。载于国际计算学习理论会议，263–277。Springer。Vinitsky, E.；Du, Y.；Parvate, K.；Jang, K.；Abbeel, P.；和 Bayen, A. 2020年。利用对抗种群进行鲁棒强化学习。*arXiv*预印本 *arXiv:2008.01825*。Wan, Y.；Naik, A.；和 Sutton, R. S. 2021年。平均奖励马尔可夫决策过程中的学习与规划。载于国际机器学习会议，10653–10662。PMLR。Wang, Y.；和 Zou, S. 2021年。模型不确定性下的在线鲁棒强化学习。载于神经信息处理系统进展（*NeurIPS*）论文集。Wang, Y.；和 Zou, S. 2022年。鲁棒强化学习的策略梯度方法。载于国际机器学习会议（*ICML*）论文集，第162卷，23484–23526。PMLR。Wei, C.-Y.；Jahromi, M. J.；Luo, H.；Sharma, H.；和 Jain, R. 2020年。无限时域平均奖励马尔可夫决策过程中的无模型强化学习。载于国际机器学习会议，10170–10180。PMLR。Wiesemann, W.；Kuhn, D.；和 Rustem, B. 2013年。鲁棒马尔可夫决策过程。《运筹学数学》，38(1)：153–183。

Xu, H.；和 Mannor, S. 2010年。分布鲁棒马尔可夫决策过程。发表于神经信息处理系统进展（*NIPS*）论文集，2505–2513。Yang, W.；Zhang, L.；和 Zhang, Z. 2021年。探究鲁棒马尔可夫决策过程的理论理解：样本复杂度与渐近性。*arXiv* 预印本 *arXiv:2105.03863*。Yu, P.；和 Xu, H. 2015年。马尔可夫决策过程中的分布鲁棒对偶问题。*IEEE* 自动控制汇刊，61(9)：2538–2543。Zhang, S.；Zhang, Z.；和 Maguluri, S. T. 2021 年。平均奖励时序差分学习和 Q-学习的有限样本分析。神经信息处理系统进展，34：1230–1242。Zhang, Y.；和 Ross, K. W. 2021年。针对平均奖励准则的同策略深度强化学习。发表于国际机器学习会议（*ICML*）论文集，12535–12545。PMLR。Zhou, Z.；Bai, Q.；Zhou, Z.；Qiu, L.；Blanchet, J.；和 Glynn, P. 2021年。分布鲁棒离线表格强化学习的有限样本遗憾界。发表于人工智能与统计国际会议（*AISTATS*）论文集，3331–3339。PMLR。

鲁棒折扣马尔可夫决策过程回顾

在本节中 在此，我们简要回顾鲁棒折扣马尔可夫决策过程的现有方法和研究结果 .

鲁棒策略评估

我们首先考虑鲁棒策略评估问题，目标是估计任意策略π的鲁棒价值函数$V_{\mathcal{P},\gamma}^{\pi}$。已有研究表明，鲁棒贝尔曼算子$\mathbf{T}_{\pi}$是一个$\gamma$-压缩映射，而鲁棒价值函数$V_{\mathcal{P},\gamma}^{\pi}$是其唯一不动点。因此，通过递归应用鲁棒贝尔曼算子，我们可以求得鲁棒折扣价值函数（Nilim 和 El Ghaoui 2004；Iyengar 2005）。

算法4：鲁棒折扣马尔可夫决策过程的策略评估

输入： π， V_0， T

1: **对于** $t = 0, 1, ..., T-1$ **执行**
2: **对于**所有 $s \in \mathcal{S}$**执行**
3: $V_{t+1}(s) \leftarrow \mathbb{E}_{\pi}[r(s, A) + \gamma\sigma_{\mathcal{P}_s^A}(V_t)]$
4: **结束循环**
5: </style>**结束循环**
6: **返回** V_T

鲁棒最优控制

鲁棒马尔可夫决策过程中的另一个重要问题是，找到使鲁棒折扣价值函数最大化的最优策略：

$$\pi^* = \arg\max_{\pi} V_{\mathcal{P},\gamma}^{\pi}. \tag{15}$$

鲁棒 值迭代方法在（Nilim 和 El Ghaoui 2004；Iyengar 2005）中提出如下： s.

算法 5：鲁棒折扣 MDP 的最优控制

输入： V_0， T

1：对于 $t = 0$**，1，...，** $T-1$**，执行 2：对于所有** $s \in \mathcal{S}$**，执行 3：** $V_{t+1}(s) \leftarrow$**最大值**$_a \{r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_t)\}$**4：结束循环 5：结束循环 6：** $\pi^*(s) \leftarrow$**最大值参数**$_a \{r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_T)\}, \forall s$**7：返回** π^*

时变模型与平稳模型之间的等价性

我们首先给出平稳策略下时变转移核模型与平稳转移核模型之间的一个等价性结果，这与鲁棒折扣马尔可夫决策过程的相关结果类似（Iyengar 2005；Nilim 和 El Ghaoui 2004）。这一结果将在后续证明中使用。

回顾式 (4) 和式 (5) 中鲁棒折扣价值函数和最坏情况平均奖励的定义，最坏情况是相对于 $\kappa = (\mathrm{P}_0, \mathrm{P}_1 ...) \in \bigotimes_{t\geq 0} \mathcal{P}$ 而取的，因此，每个时间步的转移核都可能不同。这种模型称为时变转移核模型（见 Iyengar 2005；Nilim and ElGhaoui 2004）。另一种常用设定是，不同时间步的转移核相同，这称为平稳模型（Iyengar 2005；Nilim and El Ghaoui 2004）。本文使用以下记号来区分这两种模型。用 $\mathbb{E}_{\mathrm{P}}[\cdot]$ 表示所有时间步的转移核均为同一个 P 时的期望，即平稳模型。我们还将 $g_{\mathrm{P}}^{\pi}(s) \triangleq \lim_{n\to\infty} \mathbb{E}_{\mathrm{P},\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t \middle| S_0 = s\right]$ 和 $V^{\pi}{}_{\mathrm{P},\gamma}(s) \triangleq \mathbb{E}_{\mathrm{P},\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t \middle| S_0 = s\right]$ 分别记为平稳模型 P 下的期望平均奖励和期望折扣价值函数。用 $\mathbb{E}_{\kappa}[\cdot]$ 表示时间 t 的转移核为 P_t 时的期望，即时变模型。

对于折扣情形，Nilim 和 El Ghaoui 2004年已证明，对于平稳策略 π、任意 $\gamma \in [0,1)$ 以及任意 $s \in \mathcal{S}$,

$$V_{\mathcal{P},\gamma}^{\pi}(s) = \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\pi,\kappa}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right]$$

$$= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}} \left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s \right]. \tag{16}$$

在以下定理中，我们证明了式 (16) 对鲁棒平均奖励马尔可夫决策过程的一个类似结论：如果考虑平稳策略，那么采用时变模型的鲁棒平均奖励问题可以等价地通过平稳模型求解。S

pecifically, 我们将平稳转移核模型的最坏情况平均奖励定义如下：

$$\min_{\mathsf{P}\in\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\pi,\mathsf{P}} \left[\frac{1}{n} \sum_{t=0}^{n-1} r_t \big| S_0 = s \right]. \tag{17}$$

回顾式 (5) 中时变模型的最坏情况平均奖励。我们将证明，对于任意平稳策略，求解式 (17) 等价于求解式 (5)。

定理 9. 考虑任意平稳策略 π。那么，时变模型下的最坏情况平均奖励与平稳模型下的相同：

$$\begin{aligned} g_{\mathcal{P}}^{\pi}(s) &\triangleq \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\kappa,\pi} \left[\frac{1}{n} \sum_{t=0}^{n-1} r_t | S_0 = s \right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\mathsf{P},\pi} \left[\frac{1}{n} \sum_{t=0}^{n-1} r_t \big| S_0 = s \right]. \end{aligned} \tag{18}$$

鲁棒相对价值函数也有类似的结果：

$$\begin{aligned} V_{\mathcal{P}}^{\pi}(s) &\triangleq \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi} \left[\sum_{t=0}^{\infty} (r_t - g_{\mathcal{P}}^{\pi}) | S_0 = s \right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi} \left[\sum_{t=0}^{\infty} (r_t - g_{\mathcal{P}}^{\pi}) | S_0 = s \right]. \end{aligned} \tag{19}$$

证明。 由定理 6[2]，中的鲁棒贝尔曼方程可知

$$V_{\mathcal{P}}^{\pi}(s) + g_{\mathcal{P}}^{\pi} = \sum_a \pi(a|s) \left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi}) \right). \tag{20}$$

Deno通过 $\arg\min^{p\in\mathcal{P}_s^a}(p)^\top V_{\mathcal{P}}^{\pi} \triangleq p_s^a$[3] 求得，并记作 $\mathsf{P}^{\pi} \triangleq \{p_s^a\colon\ s\in\mathcal{S},\ \ a\in\mathcal{A}\}$。由此可得 at

$$\begin{aligned} V_{\mathcal{P}}^{\pi}(s) &= \sum_a \pi(a|s) \left(r(s,a) - g_{\mathcal{P}}^{\pi} + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi}) \right) \\ &= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \sum_a \pi(a|s) \mathbb{E}_{\mathsf{P}^{\pi}}[V_{\mathcal{P}}^{\pi}(S_1) | S_0 = s, A_0 = a] \\ &= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi}[V_{\mathcal{P}}^{\pi}(S_1) | S_0 = s] \\ &= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[\sum_a \pi(a|S_1)(r(S_1,a) - g_{\mathcal{P}}^{\pi}) | S_0 = s \right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[\sum_a \pi(a|S_1) \sigma_{\mathcal{P}_{S_1}^a}(V_{\mathcal{P}}^{\pi}) | S_0 = s \right] \\ &= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[r_1 - g_{\mathcal{P}}^{\pi} | S_0 = s \right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[\sigma_{\mathcal{P}_{S_1}^{A_1}}(V_{\mathcal{P}}^{\pi}) | S_0 = s \right] \\ &= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[r_1 - g_{\mathcal{P}}^{\pi} \big| S_0 = s \right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[(p_{S_1}^{A_1})^\top V_{\mathcal{P}}^{\pi} | S_0 = s \right] \\ &= \mathbb{E}_{\mathsf{P}^{\pi},\pi} \left[r_0 - g_{\mathcal{P}}^{\pi} + r_1 - g_{\mathcal{P}}^{\pi} | S_0 = s \right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi}[V_{\mathcal{P}}^{\pi}(S_2) | S_0 = s] \end{aligned}$$

.....

[2]Th定理 6 的证明独立于定理 9，并不依赖于此处将要证明的结果。

[3]W如果有多个极小化子，则任意选取一个。

$$= \mathbb{E}_{\mathsf{P}^\pi,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|s\right]. \tag{21}$$

根据定义，以下结论始终成立：

$$\min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right] \leq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right]. \tag{22}$$

因此，这意味着平稳转移核序列 $\kappa = (\mathsf{P}^\pi、\mathsf{P}^\pi、.....)$ 是 $V_{\mathcal{P}}^{\pi}$ 的最坏情况转移核之一。因此，可以证明公式（19）。

考虑转移核 P^π。我们分别用 $g_{\mathsf{P}^\pi}^{\pi}$ 和 $V^{\pi_{\mathsf{P}^\pi}}$ 表示其非鲁棒平均回报和非鲁棒相对价值函数。根据非鲁棒贝尔曼方程（Sutton and Barto，2018年），有

$$V_{\mathsf{P}^\pi}^{\pi}(s) = \sum_a \pi(a|s)(r(s,a) - g_{\mathsf{P}^\pi}^{\pi}) + \mathbb{E}_{\mathsf{P}^\pi,\pi}[V_{\mathsf{P}^\pi}^{\pi}(S_1)|s]. \tag{23}$$

另一方面，鲁棒贝尔曼方程表明

$$V_{\mathcal{P}}^{\pi}(s) = V_{\mathsf{P}^\pi}^{\pi}(s) = \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^\pi,\pi}[V_{\mathsf{P}^\pi}^{\pi}(S_1)|s]. \tag{24}$$

因此，这两个方程意味着 $g_{\mathcal{P}}^{\pi} = g_{\mathsf{P}^\pi}^{\pi}$，，进而平稳核 $(\mathsf{P}^\pi、\mathsf{P}^\pi、.....)$ 也是时变设定下鲁棒平均回报的最坏情况核。这就证明了公式（18）。 □

定理 1 的证明

在证明中，除非另有说明，我们用 $\|v\|$ 表示向量 v的 l_∞ 范数；对于矩阵 A，我们用 $\|A\|$ 表示其由 l_∞ 范数诱导的矩阵范数，即 $\|A\| = \sup_{x\in\mathbb{R}^d} \frac{\|Ax\|_\infty}{\|x\|_\infty}$。

引理 1. *[(Puterman 1994)*中的定理 *8.2.3]* 对于任意P， γ, π,

$$V_{\mathsf{P},\gamma}^{\pi} = \frac{1}{1-\gamma} g_{\mathsf{P}}^{\pi} + h_{\mathsf{P}}^{\pi} + f_{\mathsf{P}}^{\pi}(\gamma), \tag{25}$$

其中 $h_{\mathsf{P}}^{\pi} = H_{\mathsf{P}}^{\pi} r_\pi$，且 $f_{\mathsf{P}}^{\pi}(\gamma) = \frac{1}{\gamma}\sum_{n=1}^{\infty}(-1)^n \binom{1-\gamma}{\gamma} (H_{\mathsf{P}}^{\pi})^{n+1} r_\pi$。

根据 (Puterman 1994) 中的命题 8.4.6，我们可以证明以下引理。

引理 2. H_{P}^{π} 在 $\Pi \times \mathcal{P}$ 上连续。若 Π 和 $\mathcal{P}$ 是紧致的，则 $\|H_{\mathsf{P}}^{\pi}\|$ 在 $\Pi \times \mathcal{P}$ 上一致有界，即存在常数 h，使得对于任意 π, P，都有 $\|H_{\mathsf{P}}^{\pi}\| \leq h$。

为简便起见，记为

$$S_\infty^{\pi}(\mathsf{P},\gamma) \triangleq \frac{1}{\gamma}\sum_{n=1}^{\infty}(-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H_{\mathsf{P}}^{\pi})^{n+1} r_\pi,$$
$$S_N^{\pi}(\mathsf{P},\gamma) \triangleq \frac{1}{\gamma}\sum_{n=1}^{N}(-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H_{\mathsf{P}}^{\pi})^{n+1} r_\pi. \tag{26}$$

显然，对于任意给定的 π，P， $S_\infty^{\pi}(\mathsf{P}\ \gamma) = f_{\mathsf{P}}^{\pi}(\gamma)$，且 $\lim_{N\to\infty} S_N^{\pi}(\mathsf{P}\ \gamma) = S_\infty^{\pi}(\mathsf{P}\ \gamma)$。

引理 3. 存在 $\delta \in (0,1)$，使得

$$\lim_{N\to\infty} S_N^{\pi}(\mathsf{P},\gamma) = S_\infty^{\pi}(\mathsf{P},\gamma) \tag{27}$$

在 $\Pi \times \mathcal{P} \times [\delta, 1]$ 上均匀分布。

证明. 注意到 $\|H_{\mathsf{P}}^{\pi}\| \leq h$，因此存在 δ，使得

$$\frac{1-\delta}{\delta} h \leq k < 1 \tag{28}$$

对于某个常数 k。那么对于任意 $\gamma \geq \delta$,

$$\frac{1-\gamma}{\gamma} h \leq \frac{1-\delta}{\delta} h \leq k. \tag{29}$$

此外，注意到

$$\left\|\frac{1}{\gamma}(-1)^n\left(\frac{1-\gamma}{\gamma}\right)^n(H_{\mathsf{P}}^{\pi})^{n+1}r\right\| \leq \frac{1}{\gamma}\left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1} \leq \frac{hk^n}{\delta} \triangleq M_n, \tag{30}$$

这是因为，对于诱导的 l_∞ 范数，$\|Ax\| \leq \|A\|\|x\|$ 和 $\|r_\pi\|_\infty \leq 1$。注意

$$\sum_{n=1}^{\infty} M_n = \frac{h}{\delta}\frac{k}{1-k}, \tag{31}$$

因此，根据 Weierstrass M 检验（Rudin 2022），$S_N^{\pi}(\mathsf{P}\ \gamma)$ 在 $\Pi \times \mathcal{P} \times [\delta, 1$ 上一致收敛到 $S_{\infty}^{\pi}(\mathsf{P}\ \gamma)\ ]$. □

引理 4. 存在一个一致常数 L，使得

$$\|S_N^{\pi}(\mathsf{P}, \gamma_1) - S_N^{\pi}(\mathsf{P}, \gamma_2)\| \leq L|\gamma_1 - \gamma_2|, \tag{32}$$

对于任意 $N, \pi, \mathsf{P}\ \gamma_1, \gamma_2 \in [\delta, 1]$。

证明。 我们首先证明 $\gamma S_N^{\pi}(\mathsf{P}\ \gamma) = \sum_{n=1}^{N}(-1)^n\left(\frac{1-\gamma}{\gamma}\right)^n(H_{\mathsf{P}}^{\pi})^{n+1}r_\pi \triangleq T_N^{\pi}(\mathsf{P}\ \gamma)$ 关于 l_∞ 范数是均匀利普希茨连续的，即，

$$\|T_N^{\pi}(\mathsf{P}, \gamma_1) - T_N^{\pi}(\mathsf{P}, \gamma_2)\| \leq l|\gamma_1 - \gamma_2|, \tag{33}$$

对于任意 N、π，$\mathsf{P}\ \gamma_1$，$\gamma_2 \in [\delta, 1]$ 以及某个常数 l。

显然，只需验证对于任意 π，N，P 或 γ，$\nabla T_N^{\pi}(\mathsf{P}\ \gamma)$ 一致有界，即可证明这一点。

首先，可以证明

$$\nabla T_N^{\pi}(\mathsf{P}, \gamma) = \sum_{n=1}^{N}(-1)^n n\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{-1}{\gamma^2}(H_{\mathsf{P}}^{\pi})^{n+1}r_\pi, \tag{34}$$

此外

$$\|\nabla T_N^{\pi}(\mathsf{P}, \gamma)\| \leq \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{1}{\gamma^2}h^{n+1} \triangleq l_N(\gamma). \tag{35}$$

注意

$$h\frac{1-\gamma}{\gamma}l_N(\gamma) = \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n}\frac{1}{\gamma^2}h^{n+2}, \tag{36}$$

那么，我们可以证明

$$\begin{aligned}
&\left(1 - h\frac{1-\gamma}{\gamma}\right)l_N(\gamma) \\
&= \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{1}{\gamma^2}h^{n+1} - \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n}\frac{1}{\gamma^2}h^{n+2} \\
&= \frac{1}{\gamma^2}h^2 - N\left(\frac{1-\gamma}{\gamma}\right)^{N}\frac{1}{\gamma^2}h^{N+2} + \sum_{n=2}^{N}\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{1}{\gamma^2}h^{n+1} \\
&\leq \frac{1}{\gamma^2}h^2 + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h} \\
&= \frac{h^2}{\gamma^2} + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h}.
\end{aligned} \tag{37}$$

因此，我们有

$$\|\nabla T_N^{\pi}(\mathsf{P}, \gamma)\| \leq l_N(\gamma) \leq \frac{1}{1-h\frac{1-\gamma}{\gamma}}\left(\frac{h^2}{\gamma^2} + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h}\right)$$

$$\leq \frac{1}{1-k}\left(\frac{h^2}{\delta^2}+\frac{h^2}{\delta^2}\frac{k}{1-k}\right), \tag{38}$$

这意味着 $\|\nabla T_N^\pi(\mathsf{P}\ \gamma)\|$ 存在一个均匀分布上界。

现在，我们有

$$\begin{aligned}&|S_N^\pi(\mathsf{P},\gamma_1)-S_N^\pi(\mathsf{P},\gamma_2)|\\&\leq \frac{|\gamma_2-\gamma_1|}{\gamma_1\gamma_2}\|T_N^\pi(\mathsf{P},\gamma_1)\|+\frac{\|T_N^\pi(\mathsf{P},\gamma_1)-T_N^\pi(\mathsf{P},\gamma_2)\|}{\gamma_2}.\end{aligned} \tag{39}$$

为证明 $\|T_N^\pi(\mathsf{P}\ \gamma)\|$ 一致有界，我们有

$$\begin{aligned}\|T_N^\pi(\mathsf{P},\gamma)\|&\leq \sum_{n=1}^{N}\left\|\left(\frac{1-\gamma}{\gamma}\right)^n(H_{\mathsf{P}}^\pi)^{n+1}r\right\|\\&\leq \sum_{n=1}^{N}\left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1}\\&\leq \sum_{n=1}^{N}k^n h\\&\leq h\frac{k}{1-k}.\end{aligned} \tag{40}$$

于是可得

$$\begin{aligned}&\|S_N^\pi(\mathsf{P},\gamma_1)-S_N^\pi(\mathsf{P},\gamma_2)\|\\&=\left\|\frac{\gamma_2-\gamma_1}{\gamma_1\gamma_2}T_N^\pi(\mathsf{P},\gamma_1)+\frac{T_N^\pi(\mathsf{P},\gamma_1)-T_N^\pi(\mathsf{P},\gamma_2)}{\gamma_2}\right\|\\&\leq\left(\frac{1}{\delta^2}h\frac{k}{1-k}+\frac{1}{\delta}\frac{1}{1-k}\left(\frac{h^2}{\delta^2}+\frac{h^2}{\delta^2}\frac{k}{1-k}\right)\right)|\gamma_1-\gamma_2|\\&\triangleq L|\gamma_1-\gamma_2|,\end{aligned} \tag{41}$$

其中，$L=\left(\frac{1}{\delta^2}h\frac{k}{1-k}+\frac{1}{\delta}\frac{1}{1-k}\left(\frac{h^2}{\delta^2}+\frac{h^2}{\delta^2}\frac{k}{1-k}\right)\right)$ 是一个普适常数，不依赖于 N，P π或 γ。 □

引理 5。 $S_\infty^\pi(\mathsf{P}\ \gamma)$ 当 $\gamma\to 1$ 时在 $\Pi\times\mathcal{P}$ 上均匀收敛。此外，对于任意 $\gamma>\delta$， $S_\infty^\pi(\mathsf{P}\ \gamma)$ 是 L-利普希茨的：对于任意 π，P 以及任意 γ_1， $\gamma_2\in(\delta,\ 1]$。

$$\|S_\infty^\pi(\mathsf{P},\gamma_1)-S_\infty^\pi(\mathsf{P},\gamma_2)\|\leq L|\gamma_1-\gamma_2|. \tag{42}$$

证明。 根据引理 3，对于任意 ϵ，存在 N_ϵ，使得对于任意 $n\geq N_\epsilon$， π，P， $\gamma>\delta$，

$$\|S_\infty^\pi(\mathsf{P},\gamma)-S_n^\pi(\mathsf{P},\gamma)\|<\epsilon. \tag{43}$$

因此，对于任意 γ_1， $\gamma_2\in(\delta,\ 1]$，

$$\begin{aligned}&\|S_\infty^\pi(\mathsf{P},\gamma_1)-S_\infty^\pi(\mathsf{P},\gamma_2)\|\\&\leq\|S_\infty^\pi(\mathsf{P},\gamma_1)-S_n^\pi(\mathsf{P},\gamma_1)\|+\|S_n^\pi(\mathsf{P},\gamma_1)-S_n^\pi(\mathsf{P},\gamma_2)\|+\|S_n^\pi(\mathsf{P},\gamma_2)-S_\infty^\pi(\mathsf{P},\gamma_2)\|\\&\leq 2\epsilon+\|S_n^\pi(\mathsf{P},\gamma_1)-S_n^\pi(\mathsf{P},\gamma_2)\|\\&\leq 2\epsilon+L|\gamma_1-\gamma_2|,\end{aligned} \tag{44}$$

其中最后一步来自引理 4。

因此，对于任意 ϵ，存在 $\omega=\max\{\delta,1-\epsilon\}$，使得对于任意 γ_1， $\gamma_2>\omega$，

$$\|S_\infty^\pi(\mathsf{P},\gamma_1)-S_\infty^\pi(\mathsf{P},\gamma_2)\|<(2+L)\epsilon, \tag{45}$$

因此，根据柯西准则，我们得出 $S_\infty^\pi(\mathsf{P}\ \gamma)$ 在 $\Pi\times\mathcal{P}$ 上均匀收敛。

另一方面，由于式 (44) 对任意 ϵ均成立，这意味着

$$\|S_\infty^\pi(\mathsf{P},\gamma_1)-S_\infty^\pi(\mathsf{P},\gamma_2)\|\leq L|\gamma_1-\gamma_2|, \tag{46}$$

至此证明完成。 □

现在我们证明定理 1。对于任意 P π，我们有

$$V^{\pi}_{\mathsf{P},\gamma} = \frac{1}{1-\gamma} g^{\pi}_{\mathsf{P}} + h^{\pi}_{\mathsf{P}} + f^{\pi}_{\mathsf{P}}(\gamma). \tag{47}$$

由此可得

$$(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}} + (1-\gamma)h^{\pi}_{\mathsf{P}} + (1-\gamma)f^{\pi}_{\mathsf{P}}(\gamma). \tag{48}$$

显然 (1 $-\gamma$)$h^{\pi}_{\mathsf{P}} \to 0$ 在 $\Pi \times \mathcal{P}$ 上一致成立，因为 $\|h^{\pi}_{\mathsf{P}}\| = \|H^{\pi}_{\mathsf{P}} r_{\pi}\| \le h$ 一致有界。接着，

$$\begin{aligned}
&\|(1-\gamma_1)f^{\pi}_{\mathsf{P}}(\gamma_1) - (1-\gamma_2)f^{\pi}_{\mathsf{P}}(\gamma_2)\| \\
&\le \|(1-\gamma_1)f^{\pi}_{\mathsf{P}}(\gamma_1) - (1-\gamma_1)f^{\pi}_{\mathsf{P}}(\gamma_2)\| + \|(1-\gamma_1)f^{\pi}_{\mathsf{P}}(\gamma_2) - (1-\gamma_2)f^{\pi}_{\mathsf{P}}(\gamma_2)\| \\
&\le (1-\gamma_1)L|\gamma_1-\gamma_2| + \|f^{\pi}_{\mathsf{P}}(\gamma_2)\||\gamma_1-\gamma_2|.
\end{aligned} \tag{49}$$

对于任意 π，P $\gamma > \delta$，

$$\begin{aligned}
\|f^{\pi}_{\mathsf{P}}(\gamma)\| &= \left\| \frac{1}{\gamma} \sum_{n=1}^{\infty} (-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H^{\pi}_{\mathsf{P}})^{n+1} r_{\pi} \right\| \\
&\le \left| \frac{1}{\gamma} \sum_{n=1}^{\infty} \left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1} \right| \\
&\le \frac{h}{\delta} \frac{1-\gamma}{\gamma} h \frac{1}{1-\frac{1-\gamma}{\gamma}h} \\
&\le \frac{h}{\delta} \frac{k}{1-k} \\
&\triangleq c_f.
\end{aligned} \tag{50}$$

因此，(1 $-\gamma$)$f^{\pi}_{\mathsf{P}}(\gamma) \to$ 在 $\Pi \times \mathcal{P}$ 上一致趋于 0，这是因为对于任意 π， $\|f^{\pi}_{\mathsf{P}}(\gamma)\|$ 一致有界 , $\gamma > \delta$, P .

于是我们得到e that $\lim_{\gamma\to 1}(1-\gamma)V^{\pi}{}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}}$ 在 $\mathcal{P} \times \Pi$ 上一致。这就完成了定理 1.

定理 2 的证明

我们首先证明一个引理，该引理允许我们交换 lim 和 max 的顺序。

引理 6. 如果函数 $f(x,y)$ 在 $\mathcal{X}$ 上当 $y \to y_0$ 时一致收敛于 $F(x)$，那么

$$\max_x \lim_{y\to y_0} f(x,y) = \lim_{y\to y_0} \max_x f(x,y). \tag{51}$$

证明。对于每个 $f(x,y)$，记最大值参数$_x$ $f(x,y)$ 为 $= x_y$，因此对任意 x,y，有 $f(x_y,y) \ge f(x,y)$。另记最大值参数$_x$ $F(x)$ 为 $= x'$。现在，由于 $f(x,y)$ 一致收敛于 $F(x)$，因此对于任意 ϵ，存在 δ'，使得 $\forall |y-y_0| < \delta'$，

$$|f(x,y) - F(x)| \le \epsilon \tag{52}$$

对任意 x。现在考虑 $|f(x_y,y) - F(x')|$对于 $|y-y_0| < \delta'$。如果 $f(x_y,y) - F(x') > 0$，那么

$$|f(x_y,y) - F(x')| = f(x_y,y) - F(x') = f(x_y,y) - F(x_y) + F(x_y) - F(x') \le \epsilon; \tag{53}$$

另一方面，如果 $f(x_y,y) - F(x') < 0$，那么

$$|f(x_y,y) - F(x')| = F(x') - f(x_y,y) = F(x') - f(x',y) + f(x',y) - f(x_y,y) \le \epsilon. \tag{54}$$

因此，我们证明了对于任意 ϵ，都存在 δ'，使得 $\forall |y-y_0| < \delta'$，

$$|f(x_y,y) - F(x')| = |\max_x f(x,y) - \max_x F(x)| \le \epsilon, \tag{55}$$

并且因此

$$\lim_{y\to y_0} \max_x f(x,y) = \max_x F(x) = \max_x \lim_{y\to y_0} f(x,y), \tag{56}$$

至此，证明完成。 □

然后，我们证明当折扣因子趋近于 1 时，鲁棒折扣价值函数均匀收敛到鲁棒平均奖励。

定理 10 （定理 2 的重述）。 鲁棒折扣价值函数在 Π 上均匀收敛到鲁棒平均奖励：

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma} = g^{\pi}_{\mathcal{P}}. \tag{57}$$

证明。 根据定理 9，对于平稳模型下的任意平稳策略 π, $g^{\pi}_{\mathcal{P}}(s) = \min_{\mathsf{P}\in\mathcal{P}} g^{\pi}_{\mathsf{P}}(s)$。因此，根据定理 1 中的一致收敛性，我们首先证明如下结论：

$$\begin{aligned} g^{\pi}_{\mathcal{P}} &= \min_{\mathsf{P}\in\mathcal{P}} g^{\pi}_{\mathsf{P}} \\ &= \min_{\mathsf{P}\in\mathcal{P}} \lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} \\ &\overset{(a)}{=} \lim_{\gamma\to 1} \min_{\mathsf{P}\in\mathcal{P}}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} \\ &= \lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma}, \end{aligned} \tag{58}$$

其中 (a) 是由引理 6 得出的。此外，注意到 $\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}}$ 在 $\Pi\times\mathcal{P}$ 上一致成立，因此 (58) 中的收敛在 Π 上也一致成立。至此，证明完成。 □

定理 3 的证明

定理 11 （定理 3 的重述）。 由算法 *1* 生成的 V_T 随着 $T\to\infty$ 增大而收敛到鲁棒平均奖励 $g^{\pi}_{\mathcal{P}}$。

Proof. 根据折扣鲁棒贝尔曼方程（Nilim 和 El Ghaoui 2004），可以证明

$$(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} = (1-\gamma_t)\sum_a \pi(a|s)(r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V^{\pi}_{\mathcal{P},\gamma_t})). \tag{59}$$

然后可以证明，对于任意 $s\in\mathcal{S}$,

$$\begin{aligned} &|V_{t+1}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \\ &= |V_{t+1}(s) - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) + (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \\ &\le |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + |V_{t+1}(s) - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s)| \\ &= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \\ &\quad + \left|\sum_a \pi(a|s)\Big((1-\gamma_t)r(s,a) + \gamma_t\sigma_{\mathcal{P}^a_s}(V_t) - ((1-\gamma_t)r(s,a) + \gamma_t\sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}))\Big)\right| \\ &= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + \left|\sum_a \pi(a|s)\Big(\gamma_t\sigma_{\mathcal{P}^a_s}(V_t) - \gamma_t\sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\Big)\right| \\ &= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + \gamma_t\left|\sum_a \pi(a|s)\Big(\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\Big)\right|. \end{aligned} \tag{60, 61}$$

若我们将其记为 $\Delta_t \triangleq \|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty$，则

$$\Delta_{t+1} \le \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \max_s\left\{\sum_a \pi(a|s)\left|\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\right|\right\}. \tag{62}$$

容易验证，$\sigma_{\mathcal{P}^a_s}(V)$ 是一个 1-利普希茨函数，因此 (62) 中的第二项可进一步界定为

$$\begin{aligned} &\sum_a \pi(a|s)\left|\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\right| \\ &\le \sum_a \pi(a|s)\|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty \\ &= \|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty, \end{aligned} \tag{63}$$

因此

$$\Delta_{t+1} \le \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t\Delta_t. \tag{64}$$

回顾一下

$$(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} = (1-\gamma_t)\min_{\mathsf{P}} V^{\pi}_{\mathsf{P},\gamma_t}. \tag{65}$$

令 $s_t^* \triangleq$ 最大值参数$_s |(1-\gamma_t)V^{\pi}_{\mathcal{P},\,\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\,\gamma_{t+1}}(s)|$。则可得

$$\|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty = |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s_t^*) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s_t^*)|. \tag{66}$$

注意，根据（Nilim and El Ghaoui 2004；Iyengar 2005），对于任意平稳策略 π，存在一个平稳模型 P，使得 $V^{\pi}_{\mathcal{P},\,\gamma}(s) = \mathbb{E}_{\mathsf{P},\,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right] \triangleq V^{\pi}{}_{\mathsf{P}\,\gamma}$。因此，在下文中，对于每个 γ_t，我们将 $V^{\pi}_{\mathcal{P},\,\gamma_t}$ 的最坏情况转移核记为 P_t。

如果 $(1-\gamma_t)V^{\pi}_{\mathcal{P},\,\gamma_t}(s_t^*) \geq (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\,\gamma_{t+1}}(s_t^*)$，那么

$$\begin{aligned}
&|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s_t^*) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s_t^*)| \\
&= \min_{\mathsf{P}}(1-\gamma_t)V^{\pi}_{\mathsf{P},\gamma_t}(s_t^*) - \min_{\mathsf{P}}(1-\gamma_{t+1})V^{\pi}_{\mathsf{P},\gamma_{t+1}}(s_t^*) \\
&= (1-\gamma_t)V^{\pi}_{\mathsf{P}_t,\gamma_t}(s_t^*) - (1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}}(s_t^*) \\
&= (1-\gamma_t)V^{\pi}_{\mathsf{P}_t,\gamma_t}(s_t^*) - (1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t}(s_t^*) + (1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t}(s_t^*) - (1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}}(s_t^*) \\
&\overset{(a)}{\leq} (1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t}(s_t^*) - (1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}}(s_t^*) \\
&\leq \|(1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}}\|_\infty,
\end{aligned} \tag{67}$$

其中，(a)由$(1-\gamma_t)V^{\pi}{}_{\mathsf{P}_t,\,\gamma_t}(s_t^*) =$最小值$_{\mathsf{P}}(1-\gamma_t)V^{\pi}{}_{\mathsf{P}\,\gamma_t}(s_t^*) \leq (1-\gamma_t)V^{\pi}{}_{\mathsf{P}_{t+1},\,\gamma_t}(s_t^*)$得到。

现在，根据引理 1，

$$(1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t} = g^{\pi}_{\mathsf{P}_{t+1}} + (1-\gamma_t)h^{\pi}_{\mathsf{P}_{t+1}} + (1-\gamma_t)f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t), \tag{68}$$

$$(1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}} = g^{\pi}_{\mathsf{P}_{t+1}} + (1-\gamma_{t+1})h^{\pi}_{\mathsf{P}_{t+1}} + (1-\gamma_{t+1})f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1}). \tag{69}$$

因此，对于任意 $\gamma_t > \delta$，式 (67) 可进一步界定为

$$\begin{aligned}
&\|(1-\gamma_t)V^{\pi}_{\mathsf{P}_{t+1},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathsf{P}_{t+1},\gamma_{t+1}}\|_\infty \\
&= \|(\gamma_{t+1}-\gamma_t)h^{\pi}_{\mathsf{P}_{t+1}} + (1-\gamma_t)f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t) - (1-\gamma_{t+1})f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1})\|_\infty \\
&\leq (\gamma_{t+1}-\gamma_t)\|h^{\pi}_{\mathsf{P}_{t+1}}\|_\infty + \|f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t) - f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1})\|_\infty + \|\gamma_{t+1}f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_t f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\overset{(a)}{\leq} h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \|\gamma_{t+1}f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_t f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\leq h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \|\gamma_{t+1}f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_{t+1}f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty + \|\gamma_{t+1}f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t) - \gamma_t f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\leq h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \gamma_{t+1}\|f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty + \|f^{\pi}_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty(\gamma_{t+1}-\gamma_t) \\
&\overset{(b)}{\leq} (h + L + \gamma_{t+1}L + \sup_{\pi,\mathsf{P},\gamma}\|f^{\pi}_{\mathsf{P}}(\gamma)\|_\infty)(\gamma_{t+1}-\gamma_t) \\
&\leq K(\gamma_{t+1}-\gamma_t),
\end{aligned} \tag{70}$$

其中，(a) 来自引理 5，适用于任意 $\gamma_t > \delta$；c_f 在 (50) 中定义，且 $K \triangleq h + 2L + c_f$ 是一个一致常数；而 (b) 来自引理 5。

类似地，当 $(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s_t^*) \leq (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s_t^*)$ 时，该不等式也成立。因此，对于任意满足 $\gamma_t > \delta$ 的t,

$$\Delta_{t+1} \leq K(\gamma_{t+1}-\gamma_t) + \gamma_t\Delta_t, \tag{71}$$

其中 K 是一个一致常数。

根据 (Tewari and Bartlett 2007) 中的引理 8，我们有 $\Delta_t \to 0$。注意，

$$\|V_t - g^{\pi}_{\mathcal{P}}\|_\infty \leq \|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty + \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - g^{\pi}_{\mathcal{P}}\|_\infty = \Delta_t + \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - g^{\pi}_{\mathcal{P}}\|_\infty. \tag{72}$$

结合定理 2，我们进一步得到

$$\lim_{t\to\infty}\|V_t - g^{\pi}_{\mathcal{P}}\|_\infty = 0, \tag{73}$$

由此完成证明。

□

定理 4 的证明

注意，最优鲁棒平均奖励定义为

$$g^*_{\mathcal{P}}(s) \triangleq \max_{\pi} g^{\pi}_{\mathcal{P}}(s). \tag{74}$$

我们进一步定义

$$V^*_{\mathcal{P},\gamma}(s) \triangleq \max_{\pi} V^{\pi}_{\mathcal{P},\gamma}(s). \tag{75}$$

定理 12 （定理 4 的重述）。 V_T 由算法 *2*生成，并收敛于最优鲁棒平均奖励 $g^*_{\mathcal{P}}$ 当$T \to \infty$。

证明。 首先，根据定理 2 中的一致收敛，可以证明

$$\lim_{t\to\infty} (1-\gamma_t) V^*_{\mathcal{P},\gamma_t} = g^*_{\mathcal{P}}. \tag{76}$$

接下来我们证明，对于任意 $s \in \mathcal{S}$,

$$\begin{aligned}
&|V_{t+1}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\le |V_{t+1}(s) - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s)| + |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\overset{(a)}{=} |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\quad + \left| \max_a \left((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V_t) \right) - \max_a \left(((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t})) \right) \right| \\
&\le |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\quad + \max_a \left| (1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V_t) - ((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t})) \right|,
\end{aligned} \tag{77}$$

其中（a）是由于最优鲁棒贝尔曼方程，最后一个不等式则源于以下事实：$|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$。

因此，式 (77) 可进一步界定为

$$\begin{aligned}
&|V_{t+1}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\le |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| + \gamma_t \max_a \left| \sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t}) \right|.
\end{aligned} \tag{78}$$

若记 $\Delta_t \triangleq \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty$，则

$$\Delta_{t+1} \le \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \max_{s.a} \left| \sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t}) \right|. \tag{79}$$

由于支撑函数 $\sigma_{\mathcal{P}^a_s}(V)$ 是 1-利普希茨的，因此可以证明，对于任意 s， a,

$$\left| \sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t}) \right| \le \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty. \tag{80}$$

因此

$$\Delta_{t+1} \le \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \Delta_t. \tag{81}$$

与定理 3 中的 (70) 类似，我们可以证明

$$\|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty \le K|\gamma_t - \gamma_{t+1}|, \tag{82}$$

并且，与 (Tewari and Bartlett 2007) 中的引理 8 类似，

$$\lim_{t\to\infty} \Delta_t = 0. \tag{83}$$

此外，请注意

$$\|V_t - g^*_{\mathcal{P}}\|_\infty \le \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty + \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - g^*_{\mathcal{P}}\|_\infty = \Delta_t + \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - g^*_{\mathcal{P}}\|_\infty, \tag{84}$$

结合公式（76）可得

$$\|V_t - g^*_{\mathcal{P}}\|_\infty \to 0, \tag{85}$$

因此，证明至此完成。

□

引理 7. 存在一个确定性最优策略，即 $\exists \pi^* \in \Pi_D$，使得 $g^{\pi^*}_{\mathcal{P}} = g^*_{\mathcal{P}} = \max_{\pi\in\Pi} g^{\pi}_{\mathcal{P}}$。

引理 7 的证明

引理 8.（重述引 *mma 7*）。存在一个确定性最优策略，即 $\exists \pi^* \in \Pi_D$，使得 $g_{\mathcal{P}}^{\pi^*} = g_{\mathcal{P}}^* = \max_{\pi \in \Pi} g_{\mathcal{P}}^{\pi}$.

证明。 假设不存在确定性最优鲁棒策略，即存在一个严格随机策略 $\pi_r \in \Pi$，使得对于任意确定性策略 $\pi \in \Pi_D$,

$$g_{\mathcal{P}}^{\pi_r} > g_{\mathcal{P}}^{\pi}. \tag{86}$$

根据定理 2，我们有

$$\lim_{\gamma \to 1}(1-\gamma)V_{\mathcal{P},\gamma}^{\pi_r} = g_{\mathcal{P}}^{\pi_r}, \tag{87}$$

$$\lim_{\gamma \to 1}(1-\gamma)V_{\mathcal{P},\gamma}^{\pi} = g_{\mathcal{P}}^{\pi}, \forall \pi \in \Pi_D. \tag{88}$$

由于确定性策略的数量有限，因此存在 $\delta < 1$，使得对于任意 $\gamma > \delta$,

$$V_{\mathcal{P},\gamma}^{\pi_r} > V_{\mathcal{P},\gamma}^{\pi}, \forall \pi \in \Pi_D. \tag{89}$$

这意味着对于 $\gamma > \delta$，随机策略 π_r 优于所有确定性策略，即

$$V_{\mathcal{P},\gamma}^{\pi_r} > \max_{\pi \in \Pi_D} V_{\mathcal{P},\gamma}^{\pi}. \tag{90}$$

然而，T Iyengar 2005 的定理 3.1 表明，存在确定性最优鲁棒策略，即,

$$\max_{\pi \in \Pi_D} V_{\mathcal{P},\gamma}^{\pi} = \max_{\pi \in \Pi} V_{\mathcal{P},\gamma}^{\pi} \geq V_{\mathcal{P},\gamma}^{\pi_r}, \tag{91}$$

这与（90）矛盾。因此可知，存在确定性最优鲁棒策略，证明完毕。 □

定理 5 的证明

定理 13（定理 5 的重述）。 存在 $0 < \delta < 1$，使得对于任意 $\gamma > \delta$，鲁棒折扣价值函数 $V_{\mathcal{P},\gamma}^*$ 的确定性最优鲁棒策略也是鲁棒平均奖励的最优策略，即,

$$V_{\mathcal{P},\gamma}^{\pi^*} = V_{\mathcal{P},\gamma}^*. \tag{92}$$

此外，当 最大值参数 $\max_{\pi \in \Pi^D} g_{\mathcal{P}}^{\pi}$ 为单元素集合时，存在唯一的 *Blackwell*最优策略。

证明。 根据引理 7，存在 $\pi^* \in \Pi^D$ 使得

$$g_{\mathcal{P}}^* = g_{\mathcal{P}}^{\pi^*}. \tag{93}$$

假设所有确定性策略的鲁棒平均奖励按降序排列:

$$g_{\mathcal{P}}^* = g_{\mathcal{P}}^{\pi_1^*} = g_{\mathcal{P}}^{\pi_2^*} = ... = g_{\mathcal{P}}^{\pi_m^*} > g_{\mathcal{P}}^{\pi_1} \geq ... \geq g_{\mathcal{P}}^{\pi_n} \tag{94}$$

对所有 π_i^*， $\pi_i \in \Pi^D$，并定义 $\Pi^* = \{\pi_i^*\colon\ i = 1,\ ...,\ m\}$。记 $d = g_{\mathcal{P}}^{\pi_i^*} - g_{\mathcal{P}}^{\pi_1}$。

根据定理 2，我们知道对于任意 $\pi \in \Pi^D$,

$$\lim_{\gamma \to 1}(1-\gamma)V_{\mathcal{P},\gamma}^{\pi} = g_{\mathcal{P}}^{\pi}. \tag{95}$$

由于集合 Π^D 是有限的，对于任意 $\epsilon < \frac{d}{2}$，存在 $\delta' < 1$，使得对于任意 $\gamma > \delta'$、 π_i^*和 π_j,

$$|(1-\gamma)V_{\mathcal{P},\gamma}^{\pi_i^*} - g_{\mathcal{P}}^*| < \epsilon, \tag{96}$$

$$|(1-\gamma)V_{\mathcal{P},\gamma}^{\pi_j} - g_{\mathcal{P}}^{\pi_j}| < \epsilon. \tag{97}$$

因此可推出

$$(1-\gamma)V_{\mathcal{P},\gamma}^{\pi_i^*} \geq (d - 2\epsilon) + (1-\gamma)V_{\mathcal{P},\gamma}^{\pi_j} > (1-\gamma)V_{\mathcal{P},\gamma}^{\pi_j}, \tag{98}$$

and

$$V_{\mathcal{P},\gamma}^{\pi_i^*} > V_{\mathcal{P},\gamma}^{\pi_j}. \tag{99}$$

注意，根据 (Iyengar 2005) 中的定理 3.1，即 $\max_{\pi \in \Pi^D} V_{\mathcal{P},\gamma}^{\pi} = V_{\mathcal{P},\gamma}^*$，对于任意 γ，都存在一个确定性策略 $\pi \in \Pi^D$，使得 $V_{\mathcal{P},\gamma}^* = V_{\mathcal{P},\gamma}^{\pi}$。结合 (99)，这意味着 $V_{\mathcal{P},\gamma}^{\pi}$ 的所有可能最优鲁棒策略都属于 $\{\pi_1^*, ...\pi_m^*\}$，即集合 Π^*。因此，存在 $\pi_j^* \in \Pi^*$，使得

$$V_{\mathcal{P},\gamma}^{\pi_j^*} = \max_{\pi \in \Pi^D} V_{\mathcal{P},\gamma}^{\pi} = V_{\mathcal{P},\gamma}^*. \tag{100}$$

对于第二部分，当鲁棒平均奖励的最优鲁棒策略唯一时，即 $\Pi^* = \{\pi^*\}$。根据上述结果，存在 δ'，使得对于任意 $\gamma > \delta'$, $V_{\mathcal{P},\gamma}^{\pi^*} > V_{\mathcal{P},\gamma}^{\pi}$对于任意 $\pi^* \neq \pi \in \Pi^D$，因此 π^* 是折扣鲁棒马尔可夫决策过程的最优策略，也是唯一的 Blackwell最优策略。

□

直接方法的结果证明

回顾

$$V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right], \tag{101}$$

其中

$$g_{\mathcal{P}}^{\pi} = \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\kappa,\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right]. \tag{102}$$

我们首先证明，鲁棒相对函数始终有限。

引理 9. 对于任意 π, $V_{\mathcal{P}}^{\pi}$ 均有限。

证明. 根据定理 9, $V_{\mathcal{P}}^{\pi} = \min_{\mathsf{P}\in\mathcal{P}} V^{\pi}{}_{\mathsf{P}} = \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\right]$. 注意到 $V_{\mathcal{P}}^{\pi}$ 可以改写为

$$\begin{aligned} V_{\mathcal{P}}^{\pi} &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}\sum_{t=0}^{n}(r_t - g_{\mathcal{P}}^{\pi})\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}\sum_{t=0}^{n}(r_t - g_{\mathsf{P}}^{\pi} + g_{\mathsf{P}}^{\pi} - g_{\mathcal{P}}^{\pi})\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi} + ng_{\mathsf{P}}^{\pi} - ng_{\mathcal{P}}^{\pi})\right], \end{aligned} \tag{103}$$

其中 $R_n = \sum_{t=0}^{n} r_t$. 注意，对于任意 $\mathsf{P} \in \mathcal{P}$ 和 n, $ng_{\mathsf{P}}^{\pi} \geq ng_{\mathcal{P}}^{\pi}$，因此

$$\lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi} + ng_{\mathsf{P}}^{\pi} - ng_{\mathcal{P}}^{\pi}) \geq \lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi}), \tag{104}$$

因此，$V_{\mathcal{P}}^{\pi}$ 的下界可推导如下，

$$\begin{aligned} V_{\mathcal{P}}^{\pi} &\geq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}}^{\pi})\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} V_{\mathsf{P}}^{\pi} \\ &= \min_{\mathsf{P}\in\mathcal{P}} H_{\mathsf{P}}^{\pi} r_{\pi}. \end{aligned} \tag{105}$$

由于以下事实，该值是有限的：H_{P}^{π} 在紧致集 $\mathcal{P}$ 上是连续的。

根据定理 9，记 $g_{\mathcal{P}}^{\pi}$ 的平稳最坏情况转移核为 P_g。通过注意到以下事实，可以得到 $V_{\mathcal{P}}^{\pi}$ 的上界：

$$\begin{aligned} V_{\mathcal{P}}^{\pi} &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}_g}^{\pi})\right] \\ &\leq \mathbb{E}_{\mathsf{P}_g,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}_g}^{\pi})\right] \\ &= V_{\mathsf{P}_g}^{\pi}, \end{aligned} \tag{106}$$

该值同样是有限的，并且 P_g 表示 $g_{\mathcal{P}}^{\pi}$的最坏情况转移核。因此，我们证明了对于任意 π， $V_{\mathcal{P}}^{\pi}$ 都是有限的，从而完成证明。□

在证明鲁棒相对价值函数定义良好之后，我们给出以下适用于平均回报鲁棒马尔可夫决策过程的鲁棒贝尔曼方程。

定理 14（重述 定理 6)。 对于任意 s 和 π，（$V_{\mathcal{P}}^{\pi}$、 $g_{\mathcal{P}}^{\pi}$）都是以下鲁棒贝尔曼方程的解 :

$$V(s) + g = \sum_{a}\pi(a|s)\left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V)\right). \tag{107}$$

证明。 根据定义，

$$V_{\mathcal{P}}^{\pi}(s) = \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right], \tag{108}$$

因此

$$\begin{aligned}
V_{\mathcal{P}}^{\pi}(s) &= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right] \\
&= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[(r_0 - g_{\mathcal{P}}^{\pi}) + \sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right] \\
&= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \left\{\sum_a \pi(a|s) r(s,a) - g_{\mathcal{P}}^{\pi} + \mathbb{E}_{\kappa,\pi}\left[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \left\{\sum_{a,s'} \pi(a|s) \mathsf{P}_{s,s'}^{a} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\right]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \min_{\mathsf{P}_0 \in \mathcal{P}} \min_{\kappa = (\mathsf{P}_1,\ldots) \in \bigotimes_{t\geq 1}\mathcal{P}} \left\{\sum_{a,s'} \pi(a|s) (\mathsf{P}_0)_{s,s'}^{a} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\right]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \min_{\mathsf{P}_0 \in \mathcal{P}} \left\{\sum_{a,s'} \pi(a|s) (\mathsf{P}_0)_{s,s'}^{a} \min_{\kappa = (\mathsf{P}_1,\ldots) \in \bigotimes_{t\geq 1}\mathcal{P}} \left\{\mathbb{E}_{\kappa,\pi}\left[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\right]\right\}\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \sum_a \pi(a|s) \sum_{s'} \min_{p_{s,s'}^{a} \in \mathcal{P}_s^a} p_{s,s'}^{a} V_{\mathcal{P}}^{\pi}(s') \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \sum_a \pi(a|s) \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi}) \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi} + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi})\right).
\end{aligned} \tag{109}$$

由此完成证明。 □

定理15。 *[定理7的重述，第 1]* 对于任意 (g, V) 若其是 $\max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) - V(s)\right\} = 0, \forall s$ 的解，则 $g = g_{\mathcal{P}}^{*}$.

证明。 在此证明中，对于两个向量 $v, w \in \mathbb{R}^n$, $v \geq w$ 表示 $v(s) \geq w(s)$ 逐项。

令 $B(g,V)(s) \triangleq \max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) - V(s)\right\}$。由于 (g,V) 是 (13) 的一个解，因此对于任意 $a \in \mathcal{A}$ 和任意 $s \in \mathcal{S}$,

$$r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) - V(s) \leq 0, \tag{110}$$

由此可知，对于任意策略 π,

$$g(s) \geq r_{\pi}(s) + \sum_a \pi(a|s) \sigma_{\mathcal{P}_s^a}(V) - V(s) \triangleq r_{\pi}(s) + \sum_a \pi(a|s)(p_s^a)^{\top} V - V(s), \tag{111}$$

其中 $r_{\pi}(s) \triangleq \sum_a \pi(a|s) r(s,a)$, $p_s^a \triangleq \arg\min_{p \in \mathcal{P}_s^a} p^{\top} V$，且 $\mathsf{P}_V = \{p_s^a \colon s \in \mathcal{S}, a \in \mathcal{A}\}$。我们还将由 π 和 P_V 诱导的状态转移矩阵记为 P_V^{π}。

使用这些记号并重写公式（111），可得

$$g\mathbb{1} \geq r_{\pi} + (\mathsf{P}_V^{\pi} - I)V. \tag{112}$$

由于公式（112）中的不等式逐项成立，P_V^{π} 的所有元素均为正，因此将公式（112）两边乘以 P_V^{π}，可得

$$g\mathbb{1} = g\mathsf{P}_V^{\pi}\mathbb{1} \geq \mathsf{P}_V^{\pi} r_{\pi} + \mathsf{P}_V^{\pi}(\mathsf{P}_V^{\pi} - I)V. \tag{113}$$

将公式（113）的两边乘以 P_V^{π}，并反复进行上述操作，可得

$$g\mathbb{1} \geq (\mathsf{P}_V^{\pi})^2 r_{\pi} + (\mathsf{P}_V^{\pi})^2(\mathsf{P}_V^{\pi} - I)V, \tag{114}$$

$$\vdots \qquad \vdots \tag{115}$$

$$g\mathbb{1} \geq (\mathsf{P}_V^\pi)^{n-1} r_\pi + (\mathsf{P}_V^\pi)^{n-1}(\mathsf{P}_V^\pi - I)V. \tag{116}$$

将从公式（112）到公式（116）的这些不等式相加，可得

$$ng\mathbb{1} \geq (I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1}) r_\pi + (I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})(\mathsf{P}_V^\pi - I)V, \tag{117}$$

由此可得

$$\begin{aligned} g\mathbb{1} &\geq \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1}) r_\pi + \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})(\mathsf{P}_V^\pi - I)V \\ &= \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1}) r_\pi + \frac{1}{n}((\mathsf{P}_V^\pi)^n - I)V. \end{aligned} \tag{118}$$

很容易验证 $\lim_{n\to\infty} \frac{1}{n}((\mathsf{P}_V^\pi)^n - I)V = 0$，因此可得

$$\begin{aligned} g\mathbb{1} &\geq \lim_{n\to\infty} \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1}) r_\pi \\ &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathsf{P}_V^\pi, \pi}\left[\sum_{t=0}^{n} r_t\right] \\ &= g^\pi_{\mathsf{P}_V^\pi}\mathbb{1} \\ &\geq g^\pi_{\mathcal{P}}\mathbb{1}. \end{aligned} \tag{119}$$

由于公式（119）对任意策略 π 均成立，因此可得 $g \geq g^*_{\mathcal{P}}$. 另一方面，由于 $B(g,\ V) = 0$，存在一个策略 τ，使得

$$g\mathbb{1} = r_\tau + (\mathsf{P}_V^\tau - I)V, \tag{120}$$

其中，r_τ，P_V^τ 的定义与 π 类似。根据定理 9，存在一个平稳转移核 $\mathsf{P}^\tau_{\text{ave}}$，使得 $g^\tau_{\mathcal{P}} = g^\tau_{\mathsf{P}^\tau_{\text{ave}}}$。我们将由 τ 和 $\mathsf{P}^\tau_{\text{ave}}$ 诱导的状态转移矩阵记为 P^τ。由于 P_V^τ 是 V 的最坏情况转移，因此可得

$$\mathsf{P}_V^\tau V \leq \mathsf{P}^\tau V. \tag{121}$$

Thus

$$g\mathbb{1} \leq r_\tau + (\mathsf{P}^\tau - I)V. \tag{122}$$

类似地，可得

$$g\mathbb{1} \leq (\mathsf{P}^\tau)^{j-1} r_\tau + (\mathsf{P}^\tau)^{j-1}(\mathsf{P}^\tau - I)V, \tag{123}$$

对 $j = 2$，.....，n，将这些不等式相加可得

$$\begin{aligned} ng\mathbb{1} &\leq (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1}) r_\tau + (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1})(\mathsf{P}^\tau)^{n-1}(\mathsf{P}^\tau - I)V \\ &= (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1}) r_\tau + ((\mathsf{P}^\tau)^n - I)V. \end{aligned} \tag{124}$$

因此

$$g\mathbb{1} \leq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathsf{P}^\tau_{\text{ave}}, \tau}\left[\sum_{t=0}^{n} r_t\right] = g^\tau_{\mathsf{P}^\tau_{\text{ave}}}\mathbb{1} = g^\tau_{\mathcal{P}}\mathbb{1} \leq g^*_{\mathcal{P}}\mathbb{1}. \tag{125}$$

因此 $g = g^*_{\mathcal{P}}$，证明至此结束。 □

定理16（定理7第2部分的重述）。 对于任何 $(g,\ V)$ 是以下问题的一个解

$$\max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V) - V(s)\right\} = 0, \forall s, \tag{126}$$

如果我们
设

$$\pi^*(s) = \arg\max_a \left\{r(s,a) + \sigma_{\mathcal{P}^a_s}(V)\right\} \tag{127}$$

对于任意 $s \in \mathcal{S}$，则 π^* 是一个最优鲁棒策略。

证明。注意，对于任意平稳策略 π，我们将$\sigma_{\mathcal{P}^\pi}(V) \triangleq (\sum_a \pi(a|s_1)^{\sigma_{\mathcal{P}^a_{s_1}}(V)} \ldots, \sum_a \pi(a|s_{|\mathcal{S}|})^{\sigma_{\mathcal{P}^a_{s_{|\mathcal{S}|}}}(V)})$记为 $\mathbb{R}^{|\mathcal{S}|}$中的一个向量。于是，式（14）等价于

$$r_{\pi^*} + \sigma_{\mathcal{P}^{\pi^*}}(V) = \max_\pi \{r_\pi + \sigma_{\mathcal{P}^\pi}(V)\}. \tag{128}$$

因此,

$$r_{\pi^*} - g + \sigma_{\mathcal{P}^{\pi^*}}(V) - V = \max_\pi \{r_\pi - g + \sigma_{\mathcal{P}^\pi}(V) - V\}. \tag{129}$$

由于 $(g,\ V)$ 是方程 (13) 的一个解，因此

$$r_{\pi^*} - g + \sigma_{\mathcal{P}^{\pi^*}}(V) - V = 0. \tag{130}$$

根据鲁棒贝尔曼方程 (12)，$(g_{\mathcal{P}}^{\pi^*},\ V_{\mathcal{P}}^{\pi^*})$ 是方程 (130) 的一个解。因此，由定理15，$g_{\mathcal{P}}^{\pi^*} = g_{\mathcal{P}}^*$，故 π^* 是最优鲁棒策略。

□

定理 17 （定理 8 的重述）。（$w_T,\ V_t$）在算法*3*中收敛到方程(13)的一个解。

证明。我们首先将更新算子记为

$$Lv(s) \triangleq \max_a (r(s,a) + \sigma_{\mathcal{P}^a_s}(v)). \tag{131}$$

现在，考虑 $sp(Lv - Lu)$。将 $\acute{s} \triangleq$ 记为 $_s(Lv(s) - Lu(s))$ 的最大值参数，将 $\grave{s} \triangleq$ 记为 $_s(Lv(s) - Lu(s))$ 的最小值参数。另将$a_v \triangleq$ 记为 $_a(r(\acute{s},\ a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(v))$ 的最大值参数，并将 $a_u \triangleq$ 记为 $_a(r(\acute{s},\ a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(u))$ 的最大值参数。于是

$$\begin{aligned}
Lv(\acute{s}) - Lu(\acute{s}) &= \max_a (r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(v)) - \max_a (r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(u)) \\
&\triangleq r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - (r(\acute{s},a_u) + \sigma_{\mathcal{P}^{a_u}_{\acute{s}}}(u)) \\
&\le r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - (r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(u)) \\
&= \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(u) \\
&\triangleq (p_{\acute{s}}^{a_v,v})^\top v - (p_{\acute{s}}^{a_v,u})^\top u,
\end{aligned} \tag{132}$$

其中 $p^{a_v}{}_{,\acute{s}}{}^{v} = \arg\min_{p \in \mathcal{P}^{a_v}_{\acute{s}}} p^\top v$和 $p^{a_v}{}_{,\acute{s}}{}^{u} = \arg\min_{p \in \mathcal{P}^{a_v}_{\acute{s}}} p^\top u$。因此，式 (132) 可以进一步界定为

$$\begin{aligned}
&Lv(\acute{s}) - Lu(\acute{s}) \\
&\le (p_{\acute{s}}^{a_v,v})^\top v - (p_{\acute{s}}^{a_v,u})^\top u \\
&\le (p_{\acute{s}}^{a_v,u})^\top (v-u).
\end{aligned} \tag{133}$$

同理,

$$Lv(\grave{s}) - Lu(\grave{s}) \ge (p_{\grave{s}}^{a_u,v})^\top (v-u). \tag{134}$$

Thus

$$sp(Lv - Lu) \le (p_{\acute{s}}^{a_v,u})^\top (v-u) - (p_{\grave{s}}^{a_u,v})^\top (v-u). \tag{135}$$

现在记 $v-u \triangleq (x_1,\ x_2,\ \ldots\ldots,\ x_n)$、$p^{a_v}{}_{,\acute{s}}{}^{u} = (p_1,\ \ldots\ldots,\ p_n)$ 以及 $p^{a_u}{}_{,\grave{s}}{}^{v} = (q_1,\ \ldots\ldots,\ q_n)$。进一步记 $b_i \triangleq \min\{p_i, q_i\}$ 那么

$$\begin{aligned}
&\sum_{i=1}^n p_i x_i - \sum_{i=1}^n q_i x_i \\
&= \sum_{i=1}^n (p_i - b_i) x_i - \sum_{i=1}^n (q_i - b_i) x_i \\
&\le \sum_{i=1}^n (p_i - b_i) \max\{x_i\} - \sum_{i=1}^n (q_i - b_i) \min\{x_i\} \\
&= \sum_{i=1}^n (p_i - b_i) sp(x) + \left(\sum_{i=1}^n (p_i - b_i) - \sum_{i=1}^n (q_i - b_i) \right) \min\{x_i\}
\end{aligned}$$

$$= \left(1 - \sum_{i=1}^{n} b_i\right) sp(x). \tag{136}$$

因此，我们证明了

$$sp(Lv - Lu) \leq \left(1 - \sum_{i=1}^{n} b_i\right) sp(v - u). \tag{137}$$

现在，根据假设 2，并参照（Puterman 1994）中的定理 8.5.3，可以证明存在 $1 > \lambda > 0$，使得对于任意 a，u，v，

$$\sum_{i=1}^{n} b_i \geq \lambda. \tag{138}$$

此外，参照（Puterman 1994）中的定理 8.5.2，可以证明，对于某个整数 J，L 是一个 J 步压缩映射算子，即：

$$sp(L^J v - L^J u) \leq (1 - \lambda) sp(v - u). \tag{139}$$

然后可以证明，相对价值迭代会收敛到最优方程的一个解，这与非鲁棒马尔可夫决策过程在平均奖励准则下的相对价值迭代类似（参见（Puterman 1994）中的定理 8.5.7、（Sigaud 和 Buffet 2013）中的第 1.6.4 节），因此，(w_t, V_t) 会在 $\epsilon \to 0$ 时收敛到方程 (13) 的一个解。 □